

SUPREMUM NORM POSTERIOR CONTRACTION AND CREDIBLE SETS FOR NONPARAMETRIC MULTIVARIATE REGRESSION

BY WILLIAM WEIMIN YOO AND SUBHASHIS GHOSAL

Université Paris Dauphine and North Carolina State University

In the setting of nonparametric multivariate regression with unknown error variance σ^2 , we study asymptotic properties of a Bayesian method for estimating a regression function f and its mixed partial derivatives. We use a random series of tensor product of B-splines with normal basis coefficients as a prior for f , and σ is either estimated using the empirical Bayes approach or is endowed with a suitable prior in a hierarchical Bayes approach. We establish pointwise, L_2 and L_∞ -posterior contraction rates for f and its mixed partial derivatives, and show that they coincide with the minimax rates. Our results cover even the anisotropic situation, where the true regression function may have different smoothness in different directions. Using the convergence bounds, we show that pointwise, L_2 and L_∞ -credible sets for f and its mixed partial derivatives have guaranteed frequentist coverage with optimal size. New results on tensor products of B-splines are also obtained in the course.

1. Introduction. Consider the nonparametric regression model

$$(1.1) \quad Y_i = f(\mathbf{X}_i) + \varepsilon_i, \quad i = 1, \dots, n,$$

where Y_i is a response variable, \mathbf{X}_i is a d -dimensional covariate, and $\varepsilon_1, \dots, \varepsilon_n$ are independent and identically distributed (i.i.d.) as $N(0, \sigma^2)$ with unknown $0 < \sigma < \infty$. The covariates are deterministic or are sampled from some fixed distribution independent of ε_i . In both cases, each \mathbf{X}_i takes values in some rectangular region in \mathbb{R}^d , which is assumed to be $[0, 1]^d$ without loss of generality. We follow the Bayesian approach by representing f by a finite linear combination of tensor products of B-splines and endowing the coefficients with a multivariate normal prior. We consider both the empirical and the hierarchical Bayes approach for the variance σ^2 . For the latter approach, a conjugate inverse-gamma prior is particularly convenient.

We study frequentist behavior of the posterior distributions and the resulting credible sets for f and its mixed partial derivatives, in terms of

MSC 2010 subject classifications: Primary 62G08; secondary 62G05, 62G15, 62G20

Keywords and phrases: Tensor product B-splines, sup-norm posterior contraction, nonparametric multivariate regression, mixed partial derivatives, anisotropic smoothness

pointwise, L_2 and L_∞ (supremum) distances. We assume that the true regression function f_0 belongs to an anisotropic Hölder space (see Definition 2.1 below), and the errors under the true distribution are sub-Gaussian.

Posterior contraction rates for regression functions in the L_2 -norm are well studied, but results for the stronger L_∞ -norm are limited. Giné and Nickl [14] studied contraction rates in L_r -metric, $1 \leq r \leq \infty$, and obtained optimal rate using conjugacy for the Gaussian white noise model and a rate for density estimation based on a random wavelet series and Dirichlet process mixture using a testing approach. In the same context, Castillo [2] introduced techniques based on semiparametric Bernstein-von Mises (BvM) theorems to obtain optimal L_∞ -contraction rates. Hoffman et al. [17] derived adaptive optimal L_∞ -contraction rate for the white noise model and also for density estimation. Scricciolo [25] applied the techniques of [14] to obtain L_∞ -rates using Gaussian kernel mixtures prior for analytic true densities.

De Jonge and van Zanten [9] used finite random series based on tensor products of B-splines to construct a prior for nonparametric regression and derived adaptive L_2 -contraction rate for the regression function in the isotropic case. A BvM theorem for the posterior of σ is treated in [10]. Shen and Ghosal [28, 29] used tensor products of B-splines respectively for Bayesian multivariate density estimation and high dimensional density regression in the anisotropic case.

Nonparametric confidence bands for an unknown function were considered by [30, 1] and more recently by [6, 13, 5]. A Bayesian approaches the problem by constructing a credible set with a prescribed posterior probability. It is then natural to ask if the credible set has adequate frequentist coverage for large sample sizes. For parametric problems, a BvM theorem concludes that Bayesian and frequentist measures of uncertainty are nearly the same in large samples. However, for the infinite dimensional normal mean model (equivalently the Gaussian white noise model), [7, 12] observed that for many true parameters in ℓ_2 , credible regions can have inadequate coverage. Leahu [20] showed that if the prior variances are chosen very big so that the support of the prior extends beyond ℓ_2 , then coverage can be obtained. Knapik et al. [18, 19] showed that for sequences with specific smoothness, by deliberately undersmoothing the prior, coverage of credible sets may be guaranteed. Sniekers and van der Vaart [31] obtained similar results for nonparametric regression using a scaled Brownian motion prior.

Castillo and Nickl [3] showed that for the Gaussian white noise model a BvM theorem can hold in weaker topologies for some natural priors, and the resulting credible sets appropriately modified will have asymptotically the correct coverage and optimal size. A similar result for the stronger L_∞ -

norm using this weak notion of BvM theorem is considered in [4]. Adaptive L_2 -credible regions with adequate frequentist coverage are constructed using the empirical Bayes approach in [34] for the Gaussian white noise model and in [27] for the nonparametric regression model using smoothing splines. In the setting of the Gaussian white noise model, Ray [23] constructed adaptive L_2 -credible sets using a weak BvM theorem, and also adaptive L_∞ -credible band using a spike and slab prior.

In this paper we consider multivariate nonparametric regression with unknown variance parameter and study posterior contraction rates and coverage of credible sets in the pointwise, L_2 - and L_∞ -senses, for the regression function f as well as its mixed partial derivatives. Study of posterior contraction rate in L_∞ -norm is important for its natural interpretation and implications for other problems such as the convergence of the mode of a function. A L_∞ -credible band is easier to visualize than a L_2 -credible set. We assume that the smoothness of the function is given but allow anisotropy, so the smoothness level may vary with the direction. Anisotropic function has applications in estimating time-dependent spectral density of a locally stationary time series (see [22]), and variable selection (see [16]).

A prior on the regression function is constructed using a finite random series of tensor products of B-splines with normally distributed coefficients. Posterior conjugacy leads to explicit expression for the posterior distribution which is convenient for computation as well as theoretical analysis. Although wavelets are also widely used to construct random series priors, B-splines have the added advantage in that mixed partial derivatives of f are expressible in terms of lower degree B-splines. This allows posterior analysis for mixed partial derivatives of f , a topic that is largely unaddressed in the literature, except implicitly as inverse problems in the Gaussian white noise model.

The paper is organized as follows. The next section introduces notations and assumptions. Section 3 describes the prior and the resulting posterior distribution. Section 4 contains main results on pointwise and L_∞ -contraction rates of f and its mixed partial derivatives. Section 5 presents results on coverage of the corresponding credible sets. Section 6 contains a simulation study of the proposed method. Proofs are in Section 7. New results on tensor products of B-splines are presented in the Appendix.

2. Assumptions and preliminaries. We describe notations and assumptions used in this paper. Given two numerical sequences a_n and b_n , $a_n = O(b_n)$ or $a_n \lesssim b_n$ means a_n/b_n is bounded, while $a_n = o(b_n)$ or $a_n \ll b_n$ means $a_n/b_n \rightarrow 0$. Also, $a_n \asymp b_n$ means $a_n = O(b_n)$ and $b_n = O(a_n)$. For

stochastic sequence Z_n , $Z_n = O_P(a_n)$ means $P(|Z_n| \leq Ca_n) \rightarrow 1$ for some constant $C > 0$. Let $\mathbb{N} = \{1, 2, \dots\}$ and $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$.

Define $\|\mathbf{x}\|_p = (\sum_{k=1}^d |x_k|^p)^{1/p}$, $1 \leq p < \infty$, $\|\mathbf{x}\|_\infty = \max_{1 \leq k \leq d} |x_k|$, and write $\|\mathbf{x}\|$ for $\|\mathbf{x}\|_2$, the Euclidean norm. We write $\mathbf{x} \leq \mathbf{y}$ if $x_k \leq y_k$, $k = 1, \dots, d$. For an $m \times m$ matrix $\mathbf{A} = ((a_{ij}))$, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ be the smallest and largest eigenvalues, and the (r, s) matrix norm of \mathbf{A} as $\|\mathbf{A}\|_{(r,s)} = \sup\{\|\mathbf{A}\mathbf{x}\|_s : \|\mathbf{x}\|_r \leq 1\}$. In particular, $\|\mathbf{A}\|_{(2,2)} = |\lambda_{\max}(\mathbf{A})|$ and $\|\mathbf{A}\|_{(\infty,\infty)} = \max_{1 \leq i \leq m} \sum_{j=1}^m |a_{ij}|$. These norms are related by $|a_{ij}| \leq \|\mathbf{A}\|_{(2,2)} \leq \|\mathbf{A}\|_{(\infty,\infty)}$ for $1 \leq i, j \leq m$. With another matrix \mathbf{B} of the same size, $\mathbf{A} \leq \mathbf{B}$ means $\mathbf{B} - \mathbf{A}$ is non-negative definite. We denote by \mathbf{I}_m the $m \times m$ identity matrix and by $\mathbf{1}_d$ the $d \times 1$ vector of ones.

For $f : U \rightarrow \mathbb{R}$ on some bounded set $U \subseteq \mathbb{R}^d$ with interior points, let $\|f\|_p$ be the L_p -norm, and $\|f\|_\infty = \sup_{x \in U} |f(x)|$. For $\mathbf{r} = (r_1, \dots, r_d)^T \in \mathbb{N}_0^d$, let $D^{\mathbf{r}}$ be the partial derivative operator $\partial^{|\mathbf{r}|} / \partial x_1^{r_1} \dots \partial x_d^{r_d}$, where $|\mathbf{r}| = \sum_{k=1}^d r_k$. If $\mathbf{r} = \mathbf{0}$, we interpret $D^{\mathbf{0}}f \equiv f$. We say $\mathbf{Z} \sim N_J(\boldsymbol{\xi}, \boldsymbol{\Omega})$ if \mathbf{Z} has a J -dimensional normal distribution with mean vector $\boldsymbol{\xi}$ and covariance matrix $\boldsymbol{\Omega}$. For a random function $\{Z(t), t \in U\}$, write $Z \sim \text{GP}(\xi, \Omega)$ if Z is a Gaussian process with $EZ(t) = \xi(t)$ and $\text{Cov}(Z(s), Z(t)) = \Omega(s, t)$.

DEFINITION 2.1. The anisotropic Hölder space $\mathcal{H}^\alpha([0, 1]^d)$ of order $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)^T$ consists of functions $f : [0, 1]^d \rightarrow \mathbb{R}$ such that $\|f\|_{\boldsymbol{\alpha}, \infty} < \infty$, where $\|\cdot\|_{\boldsymbol{\alpha}, \infty}$ is the anisotropic Hölder norm

$$(2.1) \quad \max \left\{ \|D^{\mathbf{r}} f\|_\infty + \sum_{k=1}^d \left\| D^{(\alpha_k - r_k) \mathbf{e}_k} D^{\mathbf{r}} f \right\|_\infty : \mathbf{r} \in \mathbb{N}_0^d, \sum_{k=1}^d r_k / \alpha_k < 1 \right\}$$

and $\mathbf{e}_k \in \mathbb{R}^d$ has 1 in the k th position and zero elsewhere.

Let α^* be the harmonic mean of $(\alpha_1, \dots, \alpha_d)^T$, i.e., $\alpha^{*-1} = d^{-1} \sum_{k=1}^d \alpha_k^{-1}$. For $\mathbf{x} = (x_1, \dots, x_d)^T$, we define $\mathbf{b}_{\mathbf{J}, \mathbf{q}}(\mathbf{x}) = (B_{j_1, q_1}(x_1) \dots B_{j_d, q_d}(x_d))$, $1 \leq j_k \leq J_k, k = 1, \dots, d$ to be a collection of $J = \prod_{k=1}^d J_k$ tensor-product of B-splines, where $B_{j_k, q_k}(x_k)$ is the k th component B-spline of fixed order $q_k \geq \alpha_k$, with knot sequence $0 = t_{k,0} < t_{k,1} < \dots < t_{k,N_k} < t_{k,N_k+1} = 1$, and let $J_k = q_k + N_k$ and $\mathbf{J} = (J_1, \dots, J_d)^T$. In the prior construction the knots depend on n and N_k increases to infinity with n subject to $\prod_{k=1}^d J_k \leq n$. At each $k = 1, \dots, d$, define $\delta_{k,l} = t_{k,l} - t_{k,l-1}$ to be the one-step knot increment, and let $\Delta_k = \max_{1 \leq l \leq N_k} \delta_{k,l}$ be the mesh size. We assume that the knot sequence for each direction is quasi-uniform [Definition 6.4 of [24]], that is $\Delta_k / \min_{1 \leq l \leq N_k} \delta_{k,l} \leq C$, for some $C > 0$. This assumption is satisfied for the uniform and nested uniform partitions as special cases (Examples

6.6 and 6.7 of [24]) and we can choose a subset of knots from any given knot sequence to form a quasi-uniform sequence with $C = 3$ [Lemma 6.17 of [24]].

If the design points $\mathbf{X}_i = (X_{i1}, \dots, X_{id})^T$ for $i = 1, \dots, n$, are fixed, assume that there exists a cumulative distribution function $G(\mathbf{x})$, with positive and continuous density on $[0, 1]^d$ such that

$$(2.2) \quad \sup_{\mathbf{x} \in [0, 1]^d} |G_n(\mathbf{x}) - G(\mathbf{x})| = o\left(\prod_{k=1}^d N_k^{-1}\right),$$

where $G_n(\mathbf{x}) = n^{-1} \sum_{i=1}^n \mathbb{1}_{\prod_{k=1}^d [0, X_{ik}]}(\mathbf{x})$ is the empirical distribution of $\{\mathbf{X}_i, i = 1, \dots, n\}$, with $\mathbb{1}_U(\cdot)$ the indicator function on U .

REMARK 2.1. For example, let $n = m^d$ for some $m \in \mathbb{N}$, the discrete uniform design $\mathbf{X}_i \in \{(j-1)/(m-1) : j = 1, \dots, m\}^d$ with $i = 1, \dots, n$, satisfies (2.2) with G being the uniform distribution on $[0, 1]^d$ and $N_k \lesssim n^{\alpha^* / \{\alpha_k(2\alpha^* + d)\}}$ for $k = 1, \dots, d$.

For random design points, assume $\mathbf{X}_i \stackrel{\text{i.i.d.}}{\sim} G$ with a continuous density on $[0, 1]^d$, then (2.2) holds with probability tending to one if $N_k \lesssim n^{\alpha^* / \{\alpha_k(2\alpha^* + d)\}}$ for $k = 1, \dots, d$, and $\alpha^* > d/2$ by Donsker's theorem. In this paper, we shall prove results on posterior contraction rates and credible sets based on fixed design points; the random case can be treated by conditioning on $\mathbf{X}_i, i = 1, \dots, n$.

Let $\mathbf{B} = (\mathbf{b}_{J,q}(\mathbf{X}_1), \dots, \mathbf{b}_{J,q}(\mathbf{X}_n))^T$. Each entry of $\mathbf{B}^T \mathbf{B}$ is indexed by d -dimensional multi-indices, i.e., for $\mathbf{u} = (u_1, \dots, u_d)^T$ and $\mathbf{v} = (v_1, \dots, v_d)^T$ with $1 \leq u_k, v_k \leq J_k, k = 1, \dots, d$, the (\mathbf{u}, \mathbf{v}) th entry is $(\mathbf{B}^T \mathbf{B})_{\mathbf{u}, \mathbf{v}} = \sum_{i=1}^n \prod_{k=1}^d B_{u_k, q_k}(X_{ik}) B_{v_k, q_k}(X_{ik})$. The following generalization of matrix banding property will be useful.

DEFINITION 2.2. Let $\mathbf{A} = ((a_{\mathbf{u}, \mathbf{v}}))$ be a matrix with rows and columns indexed by d -dimensional multi-indices $\mathbf{1}_d \leq \mathbf{u}, \mathbf{v} \leq \mathbf{J}$ respectively, where arrangement of the elements are arbitrary. We say that \mathbf{A} is $\mathbf{h} = (h_1, \dots, h_d)^T$ banded if $a_{\mathbf{u}, \mathbf{v}} = 0$ whenever $|u_k - v_k| > h_k$ for some $k = 1, \dots, d$.

Given $\mathbf{X}_i = (X_{i1}, \dots, X_{id})^T$ for $i = 1, \dots, n$, such that $X_{ik} \in [t_{k,l-1}, t_{k,l}]$, only q_k adjacent basis functions $(B_{l, q_k}(X_{ik}), \dots, B_{l+q_k-1, q_k}(X_{ik}))^T$ will be nonzero for $k = 1, \dots, d$. Hence if $|u_m - v_m| > q_m$ for some $m = 1, \dots, d$, then $B_{u_m, q_m}(X_{im}) B_{v_m, q_m}(X_{im}) = 0$, and we conclude $(\mathbf{B}^T \mathbf{B})_{\mathbf{u}, \mathbf{v}} = 0$. It then follows that $\mathbf{B}^T \mathbf{B}$ is $\mathbf{q} = (q_1, \dots, q_d)^T$ -banded.

Since approximation results for anisotropic functions by linear combinations of tensor-products of B-splines assume integer smoothness (see Chapter

12, Section 3 of [24]), we assume that $\alpha \in \mathbb{N}^d$. For the isotropic case, the norm in (2.1) can be generalized (see Section 2.7.1 of [35]) and the approximation rate is obtained for all smoothness levels (Theorem 22 of Chapter XII in [8]). This allows generalization of posterior contraction results for arbitrary smoothness levels. We now describe the assumption on f_0 used in this paper.

ASSUMPTION 1. Under the true distribution P_0 , $Y_i = f_0(\mathbf{X}_i) + \varepsilon_i$, such that ε_i are i.i.d. sub-Gaussian with mean 0 and variance σ_0^2 for $i = 1, \dots, n$. Also, $f_0 \in \mathcal{H}^\alpha([0, 1]^d)$ with order $\alpha = (\alpha_1, \dots, \alpha_d)^T \in \mathbb{N}^d$. If the design points are deterministic, we assume that (2.2) holds. If the design points are random, we assume that $\alpha^* > d/2$.

Let $E_0(\cdot)$ and $\text{Var}_0(\cdot)$ be the expectation and variance operators taken with respect to P_0 . We write $\mathbf{Y} = (Y_1, \dots, Y_n)^T$, $\mathbf{X} = (\mathbf{X}_1^T, \dots, \mathbf{X}_n^T)^T$, $\mathbf{F}_0 = (f_0(\mathbf{X}_1), \dots, f_0(\mathbf{X}_n))^T$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$.

3. Prior and posterior conjugacy. We induce a prior on f by representing it as a tensor-product B-splines series, i.e., $f(\mathbf{x}) = \mathbf{b}_{\mathbf{J}, \mathbf{q}}(\mathbf{x})^T \boldsymbol{\theta}$, where $\boldsymbol{\theta} = \{\theta_{j_1, \dots, j_d} : 1 \leq j_k \leq J_k, k = 1, \dots, d\}$ are the basis coefficients. Then its $\mathbf{r} = (r_1, \dots, r_d)^T$ mixed partial derivative is

$$D^{\mathbf{r}} f(\mathbf{x}) = \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1, \dots, j_d} \prod_{k=1}^d \frac{\partial^{r_k}}{\partial x_k^{r_k}} B_{j_k, q_k}(x_k).$$

Define an operator $\mathfrak{D}_{j_k}^{r_k}$ acting on θ_{j_1, \dots, j_d} such that $\mathfrak{D}_{j_k}^0 \theta_{j_1, \dots, j_d} = \theta_{j_1, \dots, j_d}$, and for $r_k \geq 1$,

$$(3.1) \quad \mathfrak{D}_{j_k}^{r_k} \theta_{j_1, \dots, j_d} = \frac{\mathfrak{D}_{j_k}^{r_k-1} \theta_{j_1, \dots, j_{k-1}, j_k+1, j_{k+1}, \dots, j_d} - \mathfrak{D}_{j_k}^{r_k-1} \theta_{j_1, \dots, j_{k-1}, j_k, j_{k+1}, \dots, j_d}}{(t_{k, j_k} - t_{k, j_k-q_k+1})/(q_k - r_k)}.$$

Furthermore, let $\mathfrak{D}^{\mathbf{r}} \theta_{j_1, \dots, j_d} = \mathfrak{D}_{j_1}^{r_1} \cdots \mathfrak{D}_{j_d}^{r_d} \theta_{j_1, \dots, j_d}$ be the application of $\mathfrak{D}_{j_k}^{r_k}$ to θ_{j_1, \dots, j_d} for all direction $k = 1, \dots, d$. Using equations (15) and (16) of Chapter X from [8], $D^{\mathbf{r}} f(\mathbf{x})$ can be written as

$$(3.2) \quad \sum_{j_1=1}^{J_1-r_1} \cdots \sum_{j_d=1}^{J_d-r_d} \mathfrak{D}^{\mathbf{r}} \theta_{j_1, \dots, j_d} \prod_{k=1}^d B_{j_k, q_k-r_k}(x_k) = \mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{x})^T \mathbf{W}_{\mathbf{r}} \boldsymbol{\theta},$$

where $\mathbf{W}_{\mathbf{r}}$ is a $\prod_{k=1}^d (J_k - r_k) \times \prod_{k=1}^d J_k$ matrix, with entries given by (8.1)–(8.4) in Lemma 8.2. These entries are coefficients associated with applying

the weighted finite differencing operator of (3.1) iteratively on $\boldsymbol{\theta}$ in all directions.

We represent the model in (1.1) by $\mathbf{Y} | (\mathbf{X}, \boldsymbol{\theta}, \sigma) \sim \mathcal{N}_n(\mathbf{B}\boldsymbol{\theta}, \sigma^2 \mathbf{I}_n)$. In this paper, we treat $\mathbf{J} = (J_1, \dots, J_d)^T$ as deterministic and allow it to depend on n, d and $\boldsymbol{\alpha}$. On the basis coefficients, we assign $\boldsymbol{\theta} | \sigma \sim \mathcal{N}_J(\boldsymbol{\eta}, \sigma^2 \boldsymbol{\Omega})$, where $\|\boldsymbol{\eta}\|_\infty$ is uniformly bounded. The entries of $\boldsymbol{\Omega}$ do not depend on n , and are indexed using d -dimensional multi-indices described above. We further assume that $\boldsymbol{\Omega}^{-1}$ is a $\mathbf{m} = (m_1, \dots, m_d)^T$ banded matrix with fixed \mathbf{m} . Note that $\boldsymbol{\Omega}$ depends on n only through its dimension $J \times J$. Furthermore, as $n \rightarrow \infty$, we assume that there exists constants $0 < c_1 \leq c_2 < \infty$ such that

$$(3.3) \quad c_1 \mathbf{I}_J \leq \boldsymbol{\Omega} \leq c_2 \mathbf{I}_J.$$

It follows that $D^r f | (\mathbf{Y}, \sigma) \sim \text{GP}(\mathbf{A}_r \mathbf{Y} + \mathbf{c}_r \boldsymbol{\eta}, \sigma^2 \Sigma_r)$, where \mathbf{A}_r , \mathbf{c}_r and the covariance kernel are defined for $\mathbf{x}, \mathbf{y} \in [0, 1]^d$ by

$$(3.4) \quad \mathbf{A}_r(\mathbf{x}) = \mathbf{b}_{J, q-r}(\mathbf{x})^T \mathbf{W}_r (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1} \mathbf{B}^T,$$

$$(3.5) \quad \mathbf{c}_r(\mathbf{x}) = \mathbf{b}_{J, q-r}(\mathbf{x})^T \mathbf{W}_r (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1} \boldsymbol{\Omega}^{-1},$$

$$(3.6) \quad \Sigma_r(\mathbf{x}, \mathbf{y}) = \mathbf{b}_{J, q-r}(\mathbf{x})^T \mathbf{W}_r (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1} \mathbf{W}_r^T \mathbf{b}_{J, q-r}(\mathbf{y}).$$

Since the posterior mean of $D^r f$ is an affine transformation of \mathbf{Y} , Assumption 1 implies that $\mathbf{A}_r \mathbf{Y} + \mathbf{c}_r \boldsymbol{\eta}$ is a sub-Gaussian process under P_0 . If $\mathbf{r} = \mathbf{0}$, defining $\mathbf{W}_0 = \mathbf{I}_J$, we obtain the conditional posterior distribution of f .

To deal with σ , observe that $\mathbf{Y} | \sigma \sim \mathcal{N}_n[\mathbf{B}\boldsymbol{\eta}, \sigma^2(\mathbf{B}\boldsymbol{\Omega}\mathbf{B}^T + \mathbf{I}_n)]$. Maximizing the corresponding log-likelihood with respect to σ leads to

$$(3.7) \quad \hat{\sigma}_n^2 = n^{-1}(\mathbf{Y} - \mathbf{B}\boldsymbol{\eta})^T (\mathbf{B}\boldsymbol{\Omega}\mathbf{B}^T + \mathbf{I}_n)^{-1} (\mathbf{Y} - \mathbf{B}\boldsymbol{\eta}).$$

Empirical Bayes then entails substituting the maximum likelihood estimator $\hat{\sigma}_n$ for σ in the conditional posterior of $D^r f$, i.e.,

$$(3.8) \quad \Pi(D^r f | \mathbf{Y}, \sigma) |_{\sigma=\hat{\sigma}_n} = \Pi_{\hat{\sigma}_n}(D^r f | \mathbf{Y}) \sim \text{GP}(\mathbf{A}_r \mathbf{Y} + \mathbf{c}_r \boldsymbol{\eta}, \hat{\sigma}_n^2 \Sigma_r).$$

In a hierarchical Bayes approach, we further endow σ with a continuous and positive prior density. A conjugate inverse-gamma (IG) prior $\sigma^2 \sim \text{IG}(\beta_1/2, \beta_2/2)$, with hyperparameters $\beta_1 > 4$ and $\beta_2 > 0$ is particularly convenient for both computation and theoretical analysis since by direct calculations, the posterior of σ^2 is

$$(3.9) \quad \sigma^2 | \mathbf{Y} \sim \text{IG}((\beta_1 + n)/2, (\beta_2 + n\hat{\sigma}_n^2)/2).$$

Under the quasi-uniformity of the knots and (2.2), Lemma 8.9 concludes that there exist constants $0 < C_1 \leq C_2 < \infty$ such that

$$(3.10) \quad C_1 n \left(\prod_{k=1}^d J_k^{-1} \right) \mathbf{I}_J \leq \mathbf{B}^T \mathbf{B} \leq C_2 n \left(\prod_{k=1}^d J_k^{-1} \right) \mathbf{I}_J.$$

In particular, $\|\mathbf{B}^T \mathbf{B}\|_{(2,2)} \asymp n \prod_{k=1}^d J_k^{-1}$. Combining the above with (3.3),

$$(3.11) \quad \begin{aligned} \left(C_1 n \prod_{k=1}^d J_k^{-1} + c_2^{-1} \right) &\leq \lambda_{\min} (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1}) \\ &\leq \lambda_{\max} (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1}) \leq \left(C_2 n \prod_{k=1}^d J_k^{-1} + c_1^{-1} \right). \end{aligned}$$

4. Posterior contraction rates. To establish posterior contraction rates for f and its mixed partial derivatives with unknown σ , a key step is showing that the empirical Bayes estimator for σ in the empirical Bayes approach or the posterior distribution of σ in the hierarchical Bayes approach, are consistent, uniformly for the true regression function f_0 satisfying $\|f_0\|_{\alpha,\infty} \leq R$ for any given $R > 0$.

PROPOSITION 4.1. *Let $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$, $k = 1, \dots, d$. Then for any $R > 0$, the following assertions holds uniformly for the true regression f_0 satisfying $\|f_0\|_{\alpha,\infty} \leq R$:*

- (a) *the empirical Bayes estimator $\hat{\sigma}_n$ converges to the true σ_0 at the rate $\max(n^{-1/2}, n^{-2\alpha^*/(2\alpha^*+d)})$;*
- (b) *if the inverse gamma prior $\text{IG}(\beta_1/2, \beta_2/2)$ is used on σ^2 , then the posterior for σ contracts at σ_0 at the same rate;*
- (c) *if the true distribution of the regression errors $\varepsilon_1, \dots, \varepsilon_n$ is Gaussian, then for any prior on σ with positive and continuous density, the posterior distribution of σ is consistent.*

For the rest of the paper, we shall treat f and its mixed partial derivatives in a unified framework by viewing f as $D^{\mathbf{0}}f$. Then the results on posterior contraction and credible sets (Section 5) for f can be recovered by setting $\mathbf{r} = \mathbf{0}$. Since an explicit expression for the conditional posterior of $D^{\mathbf{r}}f$ given σ is available due to the normal-normal conjugacy, we derive contraction rates by directly bounding posterior probabilities of deviations from the truth uniformly for σ in a shrinking neighborhood of σ_0 , which suffices in view of the consistency of the empirical Bayes estimator or that

of the posterior distribution of σ . A decomposition of the posterior mean square error into posterior variance, variance and squared bias of the posterior mean is used for pointwise contraction, and uniformized using maximal inequalities to establish contraction with respect to the supremum distance. Contraction rates below are uniform in $\|f_0\|_{\alpha,\infty} \leq R$. We write $\epsilon_{n,\mathbf{r}} = n^{-\alpha^*\{1-\sum_{k=1}^d(r_k/\alpha_k)\}/(2\alpha^*+d)}$ and $\epsilon_{n,\mathbf{r},\infty} = (\log n/n)^{\alpha^*\{1-\sum_{k=1}^d(r_k/\alpha_k)\}/(2\alpha^*+d)}$. Observe that for $\epsilon_{n,\mathbf{r}}$ and $\epsilon_{n,\mathbf{r},\infty}$ to approach 0 as $n \rightarrow \infty$, we will need $\sum_{k=1}^d(r_k/\alpha_k) < 1$. For the hierarchical Bayes approach, we do not restrict to the inverse gamma prior for σ^2 but throughout assume that its posterior is consistent uniformly for $\|f_0\|_{\alpha,\infty} \leq R$ for any $R > 0$.

THEOREM 4.2 (Pointwise contraction). *If $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \dots, d$, then for any $\mathbf{x} \in [0, 1]^d$ and $M_n \rightarrow \infty$,*

$$\text{Empirical Bayes: } E_0 \Pi_{\hat{\sigma}_n}(|D^{\mathbf{r}} f(\mathbf{x}) - D^{\mathbf{r}} f_0(\mathbf{x})| > M_n \epsilon_{n,\mathbf{r}} | \mathbf{Y}) \rightarrow 0.$$

$$\text{Hierarchical Bayes: } E_0 \Pi(|D^{\mathbf{r}} f(\mathbf{x}) - D^{\mathbf{r}} f_0(\mathbf{x})| > M_n \epsilon_{n,\mathbf{r}} | \mathbf{Y}) \rightarrow 0.$$

REMARK 4.3. The above rate of contraction holds for the L_2 -distance as well under the same set of assumptions for both empirical and hierarchical Bayes approaches. This follows since the posterior expectation of the squared L_2 -norm can be bounded by the integral of the corresponding uniform estimates of the pointwise case obtained in the proof of Theorem 4.2.

THEOREM 4.4 (L_∞ -contraction). *If $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \dots, d$, then for any $M_n \rightarrow \infty$,*

$$\text{Empirical Bayes: } E_0 \Pi_{\hat{\sigma}_n}(\|D^{\mathbf{r}} f - D^{\mathbf{r}} f_0\|_\infty > M_n \epsilon_{n,\mathbf{r},\infty} | \mathbf{Y}) \rightarrow 0.$$

$$\text{Hierarchical Bayes: } E_0 \Pi(\|D^{\mathbf{r}} f - D^{\mathbf{r}} f_0\|_\infty > M_n \epsilon_{n,\mathbf{r},\infty} | \mathbf{Y}) \rightarrow 0.$$

Note that an extra logarithmic factor appears in the L_∞ -rate in agreement with the corresponding minimax rate for the problem (see [32, 33]). A similar result for the white noise model using a prior based on wavelet basis expansion for the signal function is given by Theorem 1 of [14] for known variance. It is interesting to note that given any notion of posterior contraction and smoothness index, the same optimal J_k , $k = 1, \dots, d$, applies to f and its mixed partial derivatives, so the Bayes procedure automatically adapts to the order of the derivative to be estimated.

5. Credible sets for f and its mixed partial derivatives. We begin by constructing pointwise credible set for $D^{\mathbf{r}} f(\mathbf{x})$ at $\mathbf{x} \in [0, 1]^d$, where $\mathbf{r} \in \mathbb{N}_0^d$ satisfies $\sum_{k=1}^d(r_k/\alpha_k) < 1$. Let $\gamma_n \in [0, 1]$ be a sequence such that $\gamma_n \rightarrow 0$

as $n \rightarrow \infty$. Define z_δ to be the $(1 - \delta)$ -quantile of a standard normal. Since $\Pi(D^r f(\mathbf{x})|\mathbf{Y}, \sigma) \sim N(\mathbf{A}_r(\mathbf{x})\mathbf{Y} + \mathbf{c}_r(\mathbf{x})\boldsymbol{\eta}, \sigma^2 \Sigma_r(\mathbf{x}, \mathbf{x}))$, we can construct a $(1 - \gamma_n)$ -pointwise credible interval for $D^r f(\mathbf{x})$ from the relation

$$\Pi(g : |g(\mathbf{x}) - \mathbf{A}_r(\mathbf{x})\mathbf{Y} - \mathbf{c}_r(\mathbf{x})\boldsymbol{\eta}| \leq z_{\gamma_n/2} \sigma \sqrt{\Sigma_r(\mathbf{x}, \mathbf{x})} | \mathbf{Y}, \sigma) = 1 - \gamma_n.$$

However, as σ is unknown, we use empirical Bayes by substituting σ by $\hat{\sigma}_n$ derived in (3.7), leading to the following empirical credible set:

$$\hat{\mathcal{C}}_{n,r,\gamma_n}(\mathbf{x}) = \{g : |g(\mathbf{x}) - \mathbf{A}_r(\mathbf{x})\mathbf{Y} - \mathbf{c}_r(\mathbf{x})\boldsymbol{\eta}| \leq z_{\gamma_n/2} \hat{\sigma}_n \sqrt{\Sigma_r(\mathbf{x}, \mathbf{x})}\}.$$

For the hierarchical Bayes approach, the resulting credible region is given by $\mathcal{C}_{n,r,\gamma_n}(\mathbf{x}) = \{g : |g(\mathbf{x}) - \mathbf{A}_r(\mathbf{x})\mathbf{Y} - \mathbf{c}_r(\mathbf{x})\boldsymbol{\eta}| \leq R_{n,r,\gamma_n}(\mathbf{x})\}$, where $R_{n,r,\gamma_n}(\mathbf{x})$ is the $(1 - \gamma_n)$ -quantile of the marginal posterior distribution of $|D^r f(\mathbf{x}) - \mathbf{A}_r(\mathbf{x})\mathbf{Y} - \mathbf{c}_r(\mathbf{x})\boldsymbol{\eta}|$ after integrating out σ with respect to its posterior distribution. If the conjugate inverse-gamma prior is used on σ^2 , then the cut-off may be expressed explicitly in terms of quantiles of a generalized t-distribution. In general, the cut-off value $R_{n,r,\gamma_n}(\mathbf{x})$ may be found by posterior sampling: generate σ from its marginal posterior distribution and $D^r f | (\mathbf{Y}, \sigma) \sim \text{GP}(\mathbf{A}_r \mathbf{Y} + \mathbf{c}_r \boldsymbol{\eta}, \sigma^2 \Sigma_r)$.

THEOREM 5.1 (Pointwise credible intervals). *If $J_k \asymp n^{\alpha^* / \{\alpha_k(2\alpha^* + d)\}}$, $k = 1, \dots, d$, then for $\gamma_n \rightarrow 0$, the coverage of $\hat{\mathcal{C}}_{n,r,\gamma_n}(\mathbf{x})$ tends to 1 and its radius is $O_{P_0}(\epsilon_{n,r} \sqrt{\log(1/\gamma_n)})$ at $\mathbf{x} \in [0, 1]^d$ uniformly on $\|f_0\|_{\alpha,\infty} \leq R$.*

If the posterior distribution of σ is consistent, then the same conclusion holds for the hierarchical Bayes credible set $\mathcal{C}_{n,r,\gamma_n}(\mathbf{x})$.

REMARK 5.2. We can also define a $(1 - \gamma_n)$ -credible set in the L_2 -norm for $D^r f$ given \mathbf{Y} and σ as the set of all functions which differ from $\mathbf{A}_r(\mathbf{x})\mathbf{Y} + \mathbf{c}_r(\mathbf{x})\boldsymbol{\eta}$ in the L_2 -norm by $\sigma h_{n,r,2,\gamma_n}$, where $h_{n,r,2,\gamma_n}$ is the $1 - \gamma_n$ quantile of the L_2 -norm of $\text{GP}(0, \Sigma_r)$. Then the empirical Bayes credible set is obtained by substituting σ by $\hat{\sigma}_n$. The hierarchical Bayes credible set is obtained by replacing $\sigma h_{n,r,2,\gamma_n}$ by the $1 - \gamma_n$ quantile of $\|D^r f - \mathbf{A}_r \mathbf{Y} - \mathbf{c}_r \boldsymbol{\eta}\|_2$. Both credible regions have asymptotic coverage 1 under the assumptions in Theorem 5.1.

The $(1 - \gamma_n)$ -empirical Bayes L_∞ -credible set for $D^r f$, can be expressed as $\{g : \|g - \mathbf{A}_r \mathbf{Y} - \mathbf{c}_r \boldsymbol{\eta}\|_\infty \leq \rho_n \hat{\sigma}_n h_{n,r,\infty,\gamma_n}\}$, where h_{n,r,∞,γ_n} is the $(1 - \gamma_n)$ -quantile of the L_∞ -norm of $\text{GP}(0, \Sigma_r)$. It turns out that in order to obtain adequate frequentist coverage, this natural credible ball needs to be slightly inflated by a factor ρ_n , leading to the inflated empirical Bayes credible region

$$(5.1) \quad \hat{\mathcal{C}}_{n,r,\infty,\gamma_n}^{\rho_n} = \{g : \|g - \mathbf{A}_r \mathbf{Y} - \mathbf{c}_r \boldsymbol{\eta}\|_\infty \leq \rho_n \hat{\sigma}_n h_{n,r,\infty,\gamma_n}\}.$$

On the other hand, unlike in the pointwise or the L_2 -credible regions, we need not make $\gamma_n \rightarrow 0$, but can allow any fixed $\gamma < 1/2$. In the hierarchical Bayes approach, we consider the analogous credible ball $\mathcal{C}_{n,r,\infty,\gamma}^{\rho_n} = \{g : \|g - \mathbf{A}_r \mathbf{Y} - \mathbf{c}_r \boldsymbol{\eta}\|_\infty \leq \rho_n R_{n,r,\infty,\gamma}\}$, where $R_{n,r,\infty,\gamma}$ stands for the $(1 - \gamma)$ -quantile of the marginal posterior distribution of $\|D^r f - \mathbf{A}_r \mathbf{Y} - \mathbf{c}_r \boldsymbol{\eta}\|_\infty$ integrating out σ with respect to its posterior distribution.

THEOREM 5.3 (L_∞ -credible region). *If $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha+d)\}}$ for $k = 1, \dots, d$, then for any $\rho_n \rightarrow \infty$ and $\gamma < 1/2$, the coverage of $\widehat{\mathcal{C}}_{n,r,\infty,\gamma}^{\rho_n}$ tends to 1 and its radius is $O_{P_0}(\epsilon_{n,r,\infty,\rho_n})$ uniformly in $\|f_0\|_{\alpha,\infty} \leq R$. Moreover, if the true distribution of the regression errors is Gaussian, then we can let $\rho_n = \rho$ for some sufficiently large constant $\rho > 0$.*

If the posterior distribution of σ is consistent, then the same conclusion holds for the hierarchical Bayes L_∞ -credible ball $\mathcal{C}_{n,r,\infty,\gamma}^{\rho_n}$.

REMARK 5.4. To control the size of $\widehat{\mathcal{C}}_{n,r,\infty,\gamma}^{\rho_n}$ and ensure guaranteed frequentist coverage, we can take ρ_n to be a factor slowly tending to infinity, or a sufficiently large constant for the Gaussian situation. A similar correction factor was also used by [34] in the context of adaptive L_2 -credible region.

6. Simulation. We compare finite sample performance of pointwise credible intervals and L_∞ -credible bands for f in one dimension (i.e., $d = 1$, $r = 0$) with confidence intervals and L_∞ -confidence bands proposed by Theorem 4.1 of [36]. Following [18], we consider the true function $f_0(x) = \sqrt{2} \sum_{i=1}^\infty i^{-3/2} \sin i \cos\{(i - 1/2)\pi x\}$, $x \in [0, 1]$, which has smoothness $\alpha = 1$. We observed the signal f_0 with i.i.d. $N(0, 0.1)$ errors at covariate values at $X_i = (i - 1)/(n - 1)$ for $i = 1, \dots, n$. We use cubic B-splines (i.e., $q = 4$) with uniform knot sequence, where we added 4 duplicate knots at 0 and 1. For the prior parameters, we set $\boldsymbol{\eta} = \mathbf{0}$ and $\boldsymbol{\Omega} = \mathbf{I}_J$. We construct $(1 - \gamma_n)$ -empirical credible intervals for $\gamma_n = 5/n$ with $\widehat{\sigma}_n$ computed using (3.7). The corresponding confidence regions are constructed using Theorem 4.1 of [36] based on the least squares estimator $\widehat{f}(x) = \mathbf{b}_{J,q}(x)^T \widehat{\boldsymbol{\theta}}$ for $\widehat{\boldsymbol{\theta}} = (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \mathbf{Y}$, and $\widehat{\sigma}_n^2 = (\mathbf{Y} - \mathbf{B}\widehat{\boldsymbol{\theta}})^T (\mathbf{Y} - \mathbf{B}\widehat{\boldsymbol{\theta}}) / (n - J)$. In the Bayesian context when the smoothing parameter J is to be determined from the data, it is natural to use its posterior mode. However for a fair comparison, we used leave-one-out cross validation to determine J for both methods and also observed that the posterior mode essentially chose the same values. We conduct our experiment across sample sizes $n = 100, 300, 500, 700, 1000, 2000$. For pointwise credible and confidence intervals, we report the empirical coverage based on 1000 Monte Carlo runs for each n . All simulations were carried out in R using the `bs` function from the `splines` package.

The coverage probabilities of pointwise credible and confidence intervals are shown in Figure 1. One distinguishing feature is the downward spike

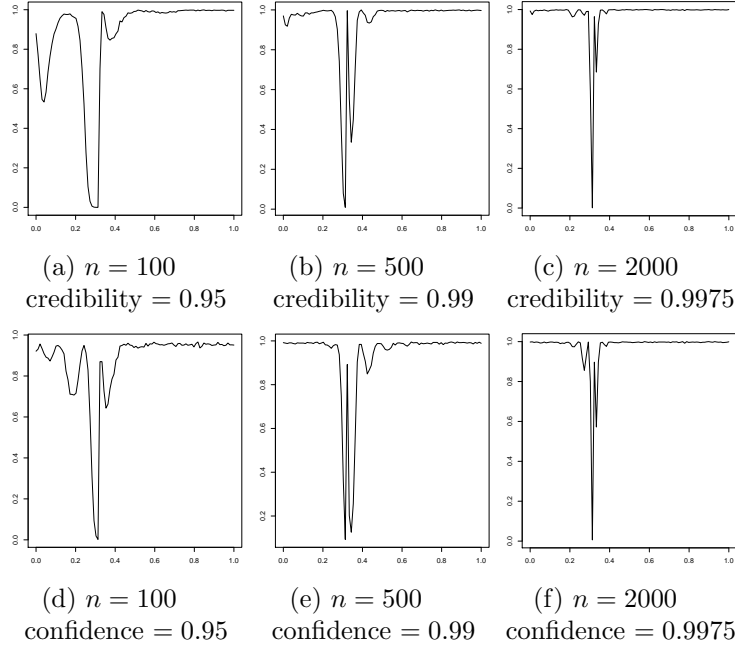


Fig 1: Pointwise coverage probabilities for credible and confidence intervals. The y -axis is the coverage probabilities and the x -axis is the covariate x .

at around the bump of f_0 at $x = 0.3$ in Figure 2, and the plots narrow down to this point as n increases. Moreover, the pointwise coverage is 0 at this point for both Bayesian and frequentist methods in all sample sizes considered. This phenomenon occurs perhaps due to the fact that the true function at $x = 0.3$ has a sharp bend but the function is much smoother elsewhere, so based on a limited sample the cross-validation method over-smooths by choosing a smaller J than ideal. Both methods yield almost the same pointwise coverage for large sample sizes, and are equivalent in quantifying uncertainty of estimating f_0 . To cover the function at all points, we consider the simultaneous (modified) credible band at the level $1 - \gamma = 0.95$, given by $(\mathbf{A}_0(x)\mathbf{Y} + \mathbf{c}_0(x)\boldsymbol{\eta}) \pm \rho\hat{\sigma}_n h_{n,0,\infty,\gamma}$.

The second assertion of Theorem 5.3 allows us to use a fixed ρ because our true errors are normally distributed which we choose as $\rho = 0.5$. To construct $(1 - \gamma)$ -asymptotic confidence band, we use Theorem 4.2 of [36]. Table 1 shows the coverage of 95% simultaneous credible and confidence bands. At

TABLE 1
95% simultaneous credible and confidence bands.

n	100	300	500	700	1000	2000
Credible band coverage	0.852	0.896	0.954	0.945	0.964	0.972
Confidence band coverage	0.972	0.948	0.963	0.978	0.985	0.986
Credible band radius	0.235	0.155	0.148	0.127	0.121	0.098
Confidence band mean radius	0.27	0.165	0.147	0.132	0.129	0.101
Confidence band max radius	0.64	0.436	0.409	0.374	0.372	0.3

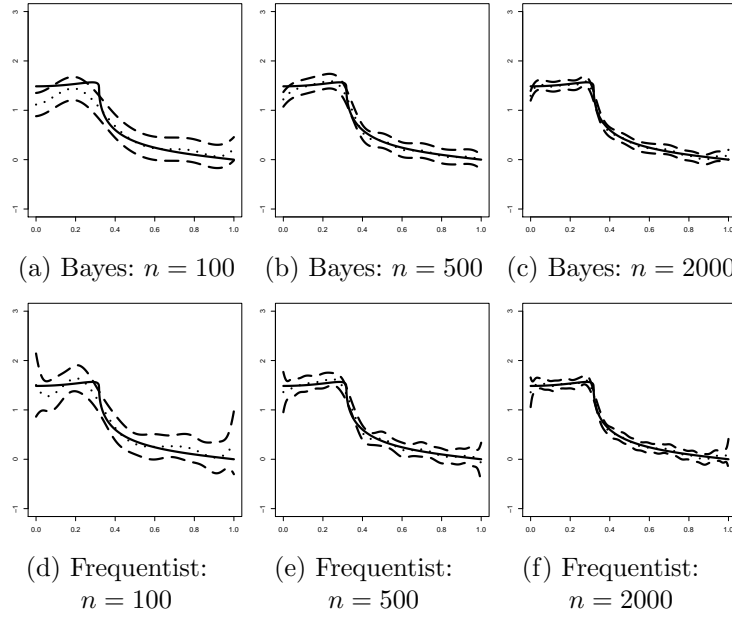


Fig 2: Dots: posterior mean (top) and \hat{f} (bottom), Solid: true function, Dashes: 95% L_∞ -credible (top) and confidence (bottom) bands.

$n = 100$, the apparent higher coverage of the confidence bands is due to the positive bias of $\tilde{\sigma}_n^2$ for small n . From $n = 300$ onward, the coverage of both credible and confidence bands steadily increase with n . The corresponding graphical representations of these bands are shown in Figure 2. The top panel corresponds to the proposed Bayesian method, where the dotted line stands for the posterior mean and dashed lines for the 95% credible band. The bottom panel corresponds to the frequentist method of [36], where the dotted line standing for the least squares estimator \hat{f} and the dashed lines for the 95% L_∞ -confidence bands. In both panels, the solid line is the true function f_0 . Observe that the credible bands have fixed length, while the

confidence bands have varying lengths. This is because the procedure of [36] is based on the supremum of the scaled absolute differences. Therefore for the latter we present both average and maximum radius. The frequentist method has larger width at the endpoints due to the fact that there are fewer observations, and this results in larger maximum radius.

7. Proofs. We shall repeatedly use the following fact about approximation power of tensor product B-splines given by (12.37) of [24].

For any $R > 0$, if $\|f_0\|_{\alpha, \infty} \leq R$, there exists a $\theta_\infty \in \mathbb{R}^J$ such that for constant $C > 0$ depending only on α, q and d , we have

$$(7.1) \quad \|\mathbf{b}_{J,q}(\cdot)^T \theta_\infty - f_0\|_\infty \leq C \sum_{k=1}^d J_k^{-\alpha_k} \left\| \frac{\partial^{\alpha_k}}{\partial x_k^{\alpha_k}} f_0 \right\|_\infty \lesssim \sum_{k=1}^d J_k^{-\alpha_k}.$$

Since $\|\mathbf{b}_{J,q}(\cdot)^T \theta_\infty\|_\infty \leq \|f_0\|_{\alpha, \infty} + C \sum_{k=1}^d J_k^{-\alpha_k} \|f_0\|_{\alpha, \infty} \lesssim R + d$,

$$(7.2) \quad \sup_{\|f_0\|_{\alpha, \infty} \leq R} \|\theta_\infty\|_\infty \lesssim \sup_{\|f_0\|_{\alpha, \infty} \leq R} \|\mathbf{b}_{J,q}(\cdot)^T \theta_\infty\|_\infty = O(1),$$

by (12.25) of [24]. An extension of the approximation result for derivatives is given by the following lemma.

LEMMA 7.1. *There exists $C > 0$ depending only on α, q and d such that for $f_0 \in \mathcal{H}^\alpha([0, 1]^d)$,*

$$\|\mathbf{b}_{J,q-r}(\cdot)^T \mathbf{W}_r \theta_\infty - D^r f_0\|_\infty \leq C \left(\sum_{k=1}^d J_k^{-(\alpha_k - r_k)} \left\| D^{(\alpha_k - r_k) \mathbf{e}_k} D^r f_0 \right\|_\infty \right).$$

PROOF. Let $I_{j_1, \dots, j_d} = \prod_{k=1}^d [t_{k, j_k - q_k}, t_{k, j_k}]$. Define a bounded linear operator $Qf(\mathbf{x}) = \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} (\lambda_{j_1, \dots, j_d} f) \prod_{k=1}^d B_{j_k, q_k}(x_k)$ on $\mathcal{H}^\alpha([0, 1]^d)$, where $\lambda_{j_1, \dots, j_d} = \prod_{k=1}^d \lambda_{j_k}$ and λ_{j_k} is the dual basis of $B_{j_k, q_k}(\cdot)$, i.e., λ_{j_k} is a linear functional such that $\lambda_{i_k} B_{j_k, q_k}(\cdot) = \mathbb{1}_{\{i_k = j_k\}}(\cdot)$ for $k = 1, \dots, d$ (see Section 4.6 of [24]). Using Theorem 13.20 of [24], there exists a tensor-product Taylor's polynomial $p_{j_1, \dots, j_d}(\mathbf{x})$ such that

$$\|D^r(f_0 - p_{j_1, \dots, j_d})|_{I_{j_1, \dots, j_d}}\|_\infty \leq C \sum_{k=1}^d J_k^{-(\alpha_k - r_k)} \left\| D^{(\alpha_k - r_k) \mathbf{e}_k} D^r f_0 \right\|_{I_{j_1, \dots, j_d}}|_{I_{j_1, \dots, j_d}}|_\infty,$$

where $f|_{I_{j_1, \dots, j_d}}$ is the restriction of f onto I_{j_1, \dots, j_d} and $C > 0$ depends only on α, q and d . By equations (12.30) and (12.31) of Theorem 12.6 in [24],

$\|(D^{\mathbf{r}} f_0 - QD^{\mathbf{r}} f_0)|_{I_{j_1, \dots, j_d}}\|_{\infty}$ is bounded above by

$$\begin{aligned} & \|D^{\mathbf{r}}(f_0 - p_{j_1, \dots, j_d})|_{I_{j_1, \dots, j_d}}\|_{\infty} + \|Q(D^{\mathbf{r}} f_0 - D^{\mathbf{r}} p_{j_1, \dots, j_d})|_{I_{j_1, \dots, j_d}}\|_{\infty} \\ & \leq C \|D^{\mathbf{r}}(f_0 - p_{j_1, \dots, j_d})|_{I_{j_1, \dots, j_d}}\|_{\infty} \\ & \leq C \sum_{k=1}^d J_k^{-(\alpha_k - r_k)} \left\| D^{(\alpha_k - r_k) \mathbf{e}_k} D^{\mathbf{r}} f_0 \Big|_{I_{j_1, \dots, j_d}} \right\|_{\infty}. \end{aligned}$$

Since $QD^{\mathbf{r}} f_0 = D^{\mathbf{r}} Qf_0$, identifying $(\boldsymbol{\theta}_{\infty})_{j_1, \dots, j_d}$ from (7.1) with $\lambda_{j_1, \dots, j_d} f_0$ and applying equations (15) and (16) of Chapter X in [8], we see that $QD^{\mathbf{r}} f_0 = \mathbf{b}_{\mathbf{J}, q-\mathbf{r}}(\cdot)^T \mathbf{W}_{\mathbf{r}} \boldsymbol{\theta}_{\infty}$. Now sum both sides over $1 \leq j_k \leq J_k, k = 1, \dots, d$. \square

PROOF OF PROPOSITION 4.1. Define $\mathbf{U} = (\mathbf{B}\boldsymbol{\Omega}\mathbf{B}^T + \mathbf{I}_n)^{-1}$ and $J = \prod_{k=1}^d J_k$. By equation (33) of page 355 in [26],

$$\begin{aligned} |\mathbb{E}_0(\hat{\sigma}_n^2) - \sigma_0^2| &= |n^{-1} \sigma_0^2 \text{tr}(\mathbf{U}) - \sigma_0^2| + n^{-1} (\mathbf{F}_0 - \mathbf{B}\boldsymbol{\eta})^T \mathbf{U} (\mathbf{F}_0 - \mathbf{B}\boldsymbol{\eta}) \\ &\lesssim n^{-1} [\text{tr}(\mathbf{I}_n - \mathbf{U}) + (\mathbf{F}_0 - \mathbf{B}\boldsymbol{\theta}_{\infty})^T \mathbf{U} (\mathbf{F}_0 - \mathbf{B}\boldsymbol{\theta}_{\infty}) \\ &\quad + (\mathbf{B}\boldsymbol{\theta}_{\infty} - \mathbf{B}\boldsymbol{\eta})^T \mathbf{U} (\mathbf{B}\boldsymbol{\theta}_{\infty} - \mathbf{B}\boldsymbol{\eta})], \end{aligned} \quad (7.3)$$

where we used $(\mathbf{x} + \mathbf{y})^T \mathbf{D}(\mathbf{x} + \mathbf{y}) \leq 2\mathbf{x}^T \mathbf{D}\mathbf{x} + 2\mathbf{y}^T \mathbf{D}\mathbf{y}$ for any $\mathbf{D} \geq \mathbf{0}$. Let $\mathbf{P}_{\mathbf{B}} = \mathbf{B}(\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T$. Let \mathbf{A} be an $m \times m$ matrix, \mathbf{C} an $m \times r$ matrix, \mathbf{T} an $r \times r$ matrix, and \mathbf{W} an $r \times m$ matrix, with \mathbf{A} and \mathbf{T} invertible. Then by the binomial inverse theorem [Theorem 18.2.8 of [15]]

$$(\mathbf{A} + \mathbf{C}\mathbf{T}\mathbf{W})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1} \mathbf{C} (\mathbf{T}^{-1} + \mathbf{W} \mathbf{A}^{-1} \mathbf{C})^{-1} \mathbf{W} \mathbf{A}^{-1}. \quad (7.4)$$

Therefore, two applications of (7.4) to \mathbf{U} yield

$$(\mathbf{B}\boldsymbol{\Omega}\mathbf{B}^T + \mathbf{I}_n)^{-1} = \mathbf{I}_n - \mathbf{B}(\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1} \mathbf{B}^T = \mathbf{I}_n - \mathbf{P}_{\mathbf{B}} + \mathbf{V}, \quad (7.5)$$

where $\mathbf{V} = \mathbf{B}(\mathbf{B}^T \mathbf{B})^{-1} [\boldsymbol{\Omega} + (\mathbf{B}^T \mathbf{B})^{-1}]^{-1} (\mathbf{B}^T \mathbf{B})^{-1} \mathbf{B}^T \geq \mathbf{0}$. Hence the first term in (7.3) is

$$n^{-1} \text{tr}(\mathbf{P}_{\mathbf{B}} - \mathbf{V}) \leq n^{-1} \text{tr}(\mathbf{P}_{\mathbf{B}}) = J/n. \quad (7.6)$$

Note $\mathbf{U} \leq \mathbf{I}_n$ since $\mathbf{B}\boldsymbol{\Omega}\mathbf{B}^T \geq \mathbf{0}$, and the second term in (7.3) is bounded by

$$n^{-1} \|\mathbf{U}\|_{(2,2)} \|\mathbf{F}_0 - \mathbf{B}\boldsymbol{\theta}_{\infty}\|^2 \leq \|\mathbf{F}_0 - \mathbf{B}\boldsymbol{\theta}_{\infty}\|_{\infty}^2 \lesssim \sum_{k=1}^d J_k^{-2\alpha_k}, \quad (7.7)$$

in view of (7.1). By (7.5) and $(\mathbf{I} - \mathbf{P}_{\mathbf{B}})\mathbf{B} = \mathbf{0}$, the last term in (7.3) is $n^{-1}(\boldsymbol{\theta}_{\infty} - \boldsymbol{\eta})^T [\boldsymbol{\Omega} + (\mathbf{B}^T \mathbf{B})^{-1}]^{-1} (\boldsymbol{\theta}_{\infty} - \boldsymbol{\eta})$, which is bounded above by

$$n^{-1} (c_1 + C_2^{-1} J/n)^{-1} J \|\boldsymbol{\theta}_{\infty} - \boldsymbol{\eta}\|_{\infty}^2 \lesssim J/n, \quad (7.8)$$

where we used (3.3) and (3.10) to bound the maximum eigenvalue of $[\mathbf{\Omega} + (\mathbf{B}^T \mathbf{B})^{-1}]^{-1}$. By (7.2) and assumption on the prior, $\|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2 = O(1)$. Combining the bounds in (7.6), (7.7) and (7.8) into (7.3), we obtain $|\mathbb{E}_0(\hat{\sigma}_n^2) - \sigma_0^2| \lesssim J/n + \sum_{k=1}^d J_k^{-2\alpha_k}$.

Let $\mathbf{Y} = \mathbf{F}_0 + \boldsymbol{\varepsilon}$ and write $n\hat{\sigma}_n^2 = (\mathbf{F}_0 - \mathbf{B}\boldsymbol{\eta})^T \mathbf{U}(\mathbf{F}_0 - \mathbf{B}\boldsymbol{\eta}) + 2(\mathbf{F}_0 - \mathbf{B}\boldsymbol{\eta})^T \mathbf{U}\boldsymbol{\varepsilon} + \boldsymbol{\varepsilon}^T \mathbf{U}\boldsymbol{\varepsilon}$. Using the fact $\text{Var}(T_1 + T_2) \leq 2\text{Var}(T_1) + 2\text{Var}(T_2)$, it follows that $\text{Var}_0(\hat{\sigma}_n^2)$ is bounded up to a constant multiple by

$$(7.9) \quad n^{-2}[(\mathbf{F}_0 - \mathbf{B}\boldsymbol{\theta}_\infty)^T \mathbf{U}^2 (\mathbf{F}_0 - \mathbf{B}\boldsymbol{\theta}_\infty) + (\mathbf{B}\boldsymbol{\theta}_\infty - \mathbf{B}\boldsymbol{\eta})^T \mathbf{U}^2 (\mathbf{B}\boldsymbol{\theta}_\infty - \mathbf{B}\boldsymbol{\eta}) + \text{Var}_0(\boldsymbol{\varepsilon}^T \mathbf{U}\boldsymbol{\varepsilon})].$$

In view of (7.1) and $\mathbf{U} \leq \mathbf{I}_n$, the first term above is bounded by

$$(7.10) \quad n^{-2}\|\mathbf{U}\|_{(2,2)}^2 \|\mathbf{F}_0 - \mathbf{B}\boldsymbol{\theta}_\infty\|^2 \leq n^{-1}\|\mathbf{F}_0 - \mathbf{B}\boldsymbol{\theta}_\infty\|_\infty^2 \lesssim n^{-1} \sum_{k=1}^d J_k^{-2\alpha_k}.$$

By the idempotency of $\mathbf{I}_n - \mathbf{P}_B$ and $(\mathbf{I}_n - \mathbf{P}_B)\mathbf{B} = \mathbf{0}$, we have that $\mathbf{B}^T(\mathbf{I}_n - \mathbf{P}_B + \mathbf{V})^2 \mathbf{B}$ is

$$\begin{aligned} \mathbf{B}^T \mathbf{V}^2 \mathbf{B} &= [\mathbf{\Omega} + (\mathbf{B}^T \mathbf{B})^{-1}]^{-1} (\mathbf{B}^T \mathbf{B})^{-1} [\mathbf{\Omega} + (\mathbf{B}^T \mathbf{B})^{-1}]^{-1} \\ &\leq [\mathbf{\Omega} + (\mathbf{B}^T \mathbf{B})^{-1}]^{-1} \leq \mathbf{B}^T \mathbf{B}. \end{aligned}$$

Therefore, in view of (7.5), the second term in (7.9) is bounded by

$$(7.11) \quad (\boldsymbol{\theta}_\infty - \boldsymbol{\eta})^T \mathbf{B}^T \mathbf{B} (\boldsymbol{\theta}_\infty - \boldsymbol{\eta}) / n^2 \leq J \|\mathbf{B}^T \mathbf{B}\|_{(2,2)} \|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2 / n \lesssim n^{-1},$$

where we used (3.10) to bound $\|\mathbf{B}^T \mathbf{B}\|_{(2,2)}$, while $\|\boldsymbol{\theta}_\infty - \boldsymbol{\eta}\|_\infty^2$ is bounded using (7.2) and the assumption on the prior. By Lemma 8.10, the last term in (7.9) is $O(n^{-1})$. Combining this with the bounds established in (7.10) and (7.11) into (7.9), we obtain $\text{Var}_0(\hat{\sigma}_n^2) \lesssim n^{-1}$. If $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \dots, d$, the mean square error is

$$(7.12) \quad \mathbb{E}_0(\hat{\sigma}_n^2 - \sigma_0^2)^2 \lesssim n^{-1} + J^2 n^{-2} + \sum_{k=1}^d J_k^{-4\alpha_k} \lesssim n^{-1} + n^{-4\alpha^*/(2\alpha^*+d)},$$

which implies the first assertion.

For the assertion (b), observe that

$$\begin{aligned} \mathbb{E}(\sigma^2 | \mathbf{Y}) &= \beta_2(\beta_1 + n - 2)^{-1} + n(\beta_1 + n - 2)^{-1} \hat{\sigma}_n^2, \\ \text{Var}(\sigma^2 | \mathbf{Y}) &= 4(\beta_1 + n - 4)^{-1} (\beta_2(\beta_1 + n - 2)^{-1} + n(\beta_1 + n - 2)^{-1} \hat{\sigma}_n^2)^2. \end{aligned}$$

Applying Markov's inequality, the posterior for σ^2 is seen to concentrate around $\hat{\sigma}_n^2$ at the rate $n^{-1/2}$, so the assertion follows from (a).

Assertion (c) can be concluded from an anisotropic extension of the estimates obtained in the proof of Theorem 4.1 together with Theorem A.1 of [10]. Indeed the posterior contracts at the rate $n^{-\alpha^*/(2\alpha^*+d)}$, and actually at the rate $n^{-1/2}$ for $\alpha^* > d/2$ by an anisotropic extension of their Theorem 4.1. Consistency can also be approached directly from the marginal model $p_{n,\sigma}$ for \mathbf{Y} given σ , where f is integrated out, by a Schwartz-type posterior consistency argument using the test $|\hat{\sigma}_n - \sigma_0| > \epsilon$, which is consistent at the true density $p_{0,n}$ by part (a). The only departure from Schwartz's argument is that in the present case it is convenient to directly establish that for any $c > 0$, $e^{cn} \int (p_{n,\sigma}/p_{0,n}) d\Pi(\sigma) \rightarrow \infty$ in probability under $p_{0,n}$ using the consistency of $\hat{\sigma}_n$ at σ_0 . \square

We write \mathcal{U}_n for a shrinking neighborhood of σ_0 such that with probability tending to one, $\hat{\sigma}_n \in \mathcal{U}_n$ and $\Pi(\sigma \in \mathcal{U}_n | \mathbf{Y}) \rightarrow 1$. We write $D^{\mathbf{r}} \tilde{f}$ for $E(D^{\mathbf{r}} f | \mathbf{Y}) = \mathbf{A}_{\mathbf{r}} \mathbf{Y} + \mathbf{c}_{\mathbf{r}} \boldsymbol{\eta}$. Recall that $\epsilon_{n,\mathbf{r}} = n^{-\alpha^* \{1 - \sum_{k=1}^d (r_k/\alpha_k)\}/(2\alpha^*+d)}$ and $\epsilon_{n,\mathbf{r},\infty} = (\log n/n)^{\alpha^* \{1 - \sum_{k=1}^d (r_k/\alpha_k)\}/(2\alpha^*+d)}$.

PROOF OF THEOREM 4.2. Recall that at $\mathbf{x} \in [0, 1]^d$, $(D^{\mathbf{r}} f(\mathbf{x}) | \mathbf{Y}, \sigma) \sim N(D^{\mathbf{r}} \tilde{f}(\mathbf{x}), \sigma^2 \Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{x}))$, with $\Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{x})$ given in (3.6). Under P_0 , $D^{\mathbf{r}} \tilde{f}(\mathbf{x})$ is a sub-Gaussian variable with mean $\mathbf{A}_{\mathbf{r}}(\mathbf{x}) \mathbf{F}_0 + \mathbf{c}_{\mathbf{r}}(\mathbf{x}) \boldsymbol{\eta}$ and variance $\sigma_0^2 \Psi_{\mathbf{r}}(\mathbf{x}, \mathbf{x})$, where $\Psi_{\mathbf{r}}(\mathbf{x}, \mathbf{y})$ is

$$\mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{x})^T \mathbf{W}_{\mathbf{r}} (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1} \mathbf{B}^T \mathbf{B} (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1} \mathbf{W}_{\mathbf{r}}^T \mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{y}).$$

Note that the posterior variance $\sigma^2 \Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{x})$ of $D^{\mathbf{r}} f$ does not depend on \mathbf{Y} , while $D^{\mathbf{r}} \tilde{f}(\mathbf{x})$ does not depend on σ . Therefore uniformly on $\|f_0\|_{\alpha,\infty} \leq R$,

$$\begin{aligned} & E_0 \sup_{\sigma \in \mathcal{U}_n} E([D^{\mathbf{r}} f(\mathbf{x}) - D^{\mathbf{r}} f_0(\mathbf{x})]^2 | \mathbf{Y}, \sigma) \\ &= \sup_{\sigma \in \mathcal{U}_n} E([D^{\mathbf{r}} f(\mathbf{x}) - D^{\mathbf{r}} \tilde{f}(\mathbf{x})]^2 | \sigma) + E_0 [D^{\mathbf{r}} \tilde{f}(\mathbf{x}) - D^{\mathbf{r}} f_0(\mathbf{x})]^2 \\ (7.13) \quad &= \sup_{\sigma \in \mathcal{U}_n} \sigma^2 \Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{x}) + \sigma_0^2 \Psi_{\mathbf{r}}(\mathbf{x}, \mathbf{x}) + [E_0 D^{\mathbf{r}} \tilde{f}(\mathbf{x}) - D^{\mathbf{r}} f_0(\mathbf{x})]^2. \end{aligned}$$

To bound $\sigma^2 \Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{x})$, first observe that $\|\mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{x})\|^2$ is bounded by

$$(7.14) \quad \prod_{k=1}^d \max_{1 \leq j_k \leq J_k} B_{j_k, q_k - r_k}(x_k) \sum_{j_1=1}^{J_1-r_1} \cdots \sum_{j_d=1}^{J_d-r_d} B_{j_k, q_k - r_k}(x_k) \leq 1.$$

In view of (3.1), each row of $\mathbf{W}_{\mathbf{r}}$ has $\prod_{k=1}^d (r_k + 1)$ nonzero entries and each column has at most $\prod_{k=1}^d (r_k + 1)$ nonzero entries. Then by Lemmas 8.2 and

8.1, each of these nonzero entries is of the order $\prod_{k=1}^d \Delta_k^{-r_k} \asymp \prod_{k=1}^d J_k^{r_k}$. Hence, both $\|\mathbf{W}_r\|_{(\infty, \infty)}$ and $\|\mathbf{W}_r^T\|_{(\infty, \infty)}$ are $O(\prod_{k=1}^d J_k^{r_k})$. Thus,

$$(7.15) \quad \|\mathbf{W}_r^T \mathbf{W}_r\|_{(2,2)} \leq \|\mathbf{W}_r^T \mathbf{W}_r\|_{(\infty, \infty)} \lesssim \prod_{k=1}^d J_k^{2r_k}.$$

By the Cauchy-Schwarz inequality, (7.14), (7.15) and (3.11), $\sigma^2 \Sigma_r(\mathbf{x}, \mathbf{x})$ over $\sigma \in \mathcal{U}_n$ is uniformly bounded by

$$(7.16) \quad \begin{aligned} & (\sigma_0^2 + o(1)) \|\mathbf{b}_{J, q-r}(\mathbf{x})\|^2 \|\mathbf{W}_r^T \mathbf{W}_r\|_{(2,2)} \left\| (B^T B + \Omega^{-1})^{-1} \right\|_{(2,2)} \\ & \lesssim \left(C_1 n \prod_{k=1}^d J_k^{-1} + c_2^{-1} \right)^{-1} \left(\prod_{k=1}^d J_k^{2r_k} \right) \lesssim n^{-1} \prod_{k=1}^d J_k^{2r_k+1}, \end{aligned}$$

Using (3.11), (3.10), (7.14) and (7.15), the variance $\sigma_0^2 \Psi_r(\mathbf{x}, \mathbf{x})$ of $D^r f(\mathbf{x})$ is bounded by

$$(7.17) \quad \begin{aligned} & \sigma_0^2 \left\| (B^T B + \Omega^{-1})^{-1} \right\|_{(2,2)}^2 \|B^T B\|_{(2,2)} \|\mathbf{b}_{J, q-r}(\mathbf{x})\|^2 \|\mathbf{W}_r^T \mathbf{W}_r\|_{(2,2)} \\ & \lesssim \left(n^{-1} \prod_{k=1}^d J_k \right)^2 \left(n \prod_{k=1}^d J_k^{-1} \right) \left(\prod_{k=1}^d J_k^{2r_k} \right) \lesssim n^{-1} \prod_{k=1}^d J_k^{2r_k+1}. \end{aligned}$$

The last term in (7.13) is bounded as

$$\begin{aligned} & |\mathbb{E}_0 D^r \tilde{f}(\mathbf{x}) - D^r f_0(\mathbf{x})| \\ & \leq \left| \mathbf{b}_{J, q-r}(\mathbf{x})^T \mathbf{W}_r (B^T B + \Omega^{-1})^{-1} (B^T F_0 + \Omega^{-1} \eta) - \mathbf{b}_{J, q-r}(\mathbf{x})^T \mathbf{W}_r \theta_\infty \right| \\ & \quad + |D^r f_0(\mathbf{x}) - \mathbf{b}_{J, q-r}(\mathbf{x})^T \mathbf{W}_r \theta_\infty|. \end{aligned}$$

By bounding the second term using Lemma 7.1 and using $\|\mathbf{b}_{J, q-r}(\mathbf{x})\|_1 = 1$, the right hand side above, up to $O(\sum_{k=1}^d J_k^{-(\alpha_k - r_k)})$, is bounded by

$$\begin{aligned} & \left| \mathbf{b}_{J, q-r}(\mathbf{x})^T \mathbf{W}_r (B^T B + \Omega^{-1})^{-1} [B^T (F_0 - B \theta_\infty) + \Omega^{-1} (\eta - \theta_\infty)] \right| \\ & \leq \left\| (B^T B + \Omega^{-1})^{-1} \right\|_{(\infty, \infty)} \|\mathbf{W}_r\|_{(\infty, \infty)} \left\{ \|B^T (F_0 - B \theta_\infty)\|_\infty \right. \\ & \quad \left. + \|\Omega^{-1}\|_{(\infty, \infty)} (\|\theta_\infty\|_\infty + \|\eta\|_\infty) \right\}. \end{aligned}$$

Since Ω^{-1} is \mathbf{m} -banded with fixed \mathbf{m} and has uniformly bounded entries, $\|\Omega^{-1}\|_{(\infty, \infty)} = O(1)$. As $B^T B$ is \mathbf{q} -banded, Lemma 8.4 and (3.11) imply that

$\|(\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1}\|_{(\infty, \infty)} \lesssim n^{-1} \prod_{k=1}^d J_k$. By (7.15), we have $\|\mathbf{W}_r\|_{(\infty, \infty)} \lesssim \prod_{k=1}^d J_k^{r_k}$. Also, $\|\boldsymbol{\theta}_\infty\|_\infty$ and $\|\boldsymbol{\eta}\|_\infty$ are both $O(1)$ by (7.2) and the assumption on the prior. Using the non-negativity of B-splines, Lemma 8.3 and (7.1), uniformly on $\|f_0\|_{\alpha, \infty} \leq R$, we bound $\|\mathbf{B}^T(\mathbf{F}_0 - \mathbf{B}\boldsymbol{\theta}_\infty)\|_\infty$ by

$$\begin{aligned} & \max_{1 \leq j_k \leq J_k, k=1, \dots, d} \sum_{i=1}^n \prod_{k=1}^d B_{j_k, q_k}(X_{ik}) |f_0(\mathbf{X}_i) - \mathbf{b}_{\mathbf{J}, \mathbf{q}}(\mathbf{X}_i)^T \boldsymbol{\theta}_\infty| \\ & \lesssim \sum_{k=1}^d J_k^{-\alpha_k} \max_{1 \leq j_k \leq J_k, k=1, \dots, d} \sum_{i=1}^n \prod_{k=1}^d B_{j_k, q_k}(X_{ik}) \lesssim n \sum_{k=1}^d J_k^{-\alpha_k} \left(\prod_{k=1}^d J_k^{-1} \right). \end{aligned}$$

Therefore, combining the bounds obtained and squaring the bias of $D^r \tilde{f}$, we have for any $\mathbf{x} \in [0, 1]^d$ uniformly on $\|f_0\|_{\alpha, \infty} \leq R$,

$$(7.18) \quad |\mathbb{E}_0 D^r \tilde{f}(\mathbf{x}) - D^r f_0(\mathbf{x})|^2 \lesssim \prod_{k=1}^d J_k^{2r_k} \left(n^{-2} \prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha_k} \right).$$

Let $P_{n, \mathbf{r}}(\mathbf{x}) = \mathbb{E}_0 \sup_{\sigma \in \mathcal{U}_n} \mathbb{E}([D^r f(\mathbf{x}) - D^r f_0(\mathbf{x})]^2 | \mathbf{Y}, \sigma)$. Combining (7.16), (7.17) and (7.18) into (7.13),

$$\begin{aligned} \sup_{\|f_0\|_{\alpha, \infty} \leq R} P_{n, \mathbf{r}}(\mathbf{x}) & \lesssim \frac{1}{n} \prod_{k=1}^d J_k^{2r_k+1} + \prod_{k=1}^d J_k^{2r_k} \left(\frac{1}{n^2} \prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha_k} \right) \\ (7.19) \quad & \lesssim \prod_{k=1}^d J_k^{2r_k} \left(\frac{1}{n} \prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha_k} \right), \end{aligned}$$

since $\prod_{k=1}^d J_k \leq n$ by the assumption. To balance the orders of the two terms on the right, let $J_k = J^{1/\alpha_k}$ for $k = 1, \dots, d$. Then the right hand side of (7.19) reduces to $O(J^{\sum_{k=1}^d (2r_k+1)/\alpha_k} / n) + O(J^{2(\sum_{k=1}^d r_k/\alpha_k - 1)})$. They will have the same order if $J \asymp n^{\alpha^*/(2\alpha^*+d)}$, and $J_k = J^{1/\alpha_k} \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \dots, d$. Hence, $P_{n, \mathbf{r}}(\mathbf{x}) = O(\epsilon_{n, \mathbf{r}}^2)$ uniformly on $\|f_0\|_{\alpha, \infty} \leq R$, implying the first assertion.

For the hierarchical Bayes procedure, the assertion similarly follows from $\mathbb{E}_0 \Pi(|D^r f(\mathbf{x}) - D^r f_0(\mathbf{x})| > M_n \epsilon_{n, \mathbf{r}} | \mathbf{Y}) \leq M_n^{-2} \epsilon_{n, \mathbf{r}}^{-2} P_{n, \mathbf{r}}(\mathbf{x}) + \mathbb{E}_0 \Pi(\sigma \notin \mathcal{U}_n | \mathbf{Y})$. \square

PROOF OF THEOREM 4.4. Recall that $(D^r f | \mathbf{Y}, \sigma) \sim \text{GP}(D^r \tilde{f}, \sigma^2 \Sigma_r)$. Let $Z_{n, \mathbf{r}} \sim \text{GP}(0, \Sigma_r)$. Under the true distribution P_0 , $D^r \tilde{f}$ is a sub-Gaussian process with mean function $\mathbf{A}_r \mathbf{F}_0 + \mathbf{c}_r \boldsymbol{\eta}$ and covariance function $\sigma_0^2 \Psi_r$. Let

$Q_{n,r}$ be a sub-Gaussian process with mean function 0 and covariance function $\sigma_0^2 \Psi_r$. Note that $Z_{n,r}$ does not depend on \mathbf{Y} and f_0 , while $D^r \tilde{f}$ does not depend on σ . Then uniformly on $\|f_0\|_{\alpha,\infty} \leq R$,

$$\begin{aligned}
 & \mathbb{E}_0 \sup_{\sigma^2 \in \mathcal{U}_n} \mathbb{E}(\|D^r f - D^r f_0\|_\infty^2 | \mathbf{Y}, \sigma) \\
 & \lesssim \sup_{\sigma \in \mathcal{U}_n} \mathbb{E}(\|D^r f - D^r \tilde{f}\|_\infty^2 | \sigma) + \mathbb{E}_0 \|D^r \tilde{f} - D^r f_0\|_\infty^2 \\
 (7.20) \quad & \lesssim \sup_{\sigma \in \mathcal{U}_n} \sigma^2 \mathbb{E} \|Z_{n,r}\|_\infty^2 + \mathbb{E} \|Q_{n,r}\|_\infty^2 + \|\mathbf{A}_r \mathbf{F}_0 + \mathbf{c}_r \boldsymbol{\eta} - D^r f_0\|_\infty^2.
 \end{aligned}$$

Since $Q_{n,r} = \mathbf{A}_r \boldsymbol{\varepsilon}$, then by Assumption 1, $Q_{n,r}$ is sub-Gaussian with respect to the semi-metric $d(\mathbf{t}, \mathbf{s}) = \sqrt{\text{Var}(Q_{n,r}(\mathbf{t}) - Q_{n,r}(\mathbf{s}))}$. Note that $Z_{n,r}$ and $Q_{n,r}$ satisfy the condition for Lemma 8.11 by Lemma 8.6. Applying Lemma 8.11 with $p = 2$, we have for any $0 < \delta_n < 1$, $\mathbb{E} \|Z_{n,r}\|_\infty^2 \lesssim \log(1/\delta_n)(n\delta_n^2 + \frac{1}{n} \prod_{k=1}^d J_k^{2r_k+1})$ in view of (7.16). Similarly, $\mathbb{E} \|Q_{n,r}\|_\infty^2 \lesssim \log(1/\delta_n)(n\delta_n^2 + \frac{1}{n} \prod_{k=1}^d J_k^{2r_k+1})$ by (7.17). Setting $\delta_n^2 \asymp n^{-2} \prod_{k=1}^d J_k^{2r_k+1}$,

$$(7.21) \quad \mathbb{E} \|Z_{n,r}\|_\infty^2 \lesssim \frac{\log n}{n} \prod_{k=1}^d J_k^{2r_k+1}, \quad \mathbb{E} \|Q_{n,r}\|_\infty^2 \lesssim \frac{\log n}{n} \prod_{k=1}^d J_k^{2r_k+1}.$$

Since the bound for (7.18) is uniform for $\mathbf{x} \in [0, 1]^d$ and $\|f_0\|_{\alpha,\infty} \leq R$,

$$(7.22) \quad \|\mathbf{A}_r \mathbf{F}_0 + \mathbf{c}_r \boldsymbol{\eta} - D^r f_0\|_\infty^2 \lesssim \prod_{k=1}^d J_k^{2r_k} \left(n^{-2} \prod_{k=1}^d J_k^2 + \sum_{k=1}^d J_k^{-2\alpha_k} \right).$$

Combining (7.21) and (7.22) with (7.20), uniformly on $\|f_0\|_{\alpha,\infty} \leq R$,

$$\mathbb{E}_0 \sup_{\sigma \in \mathcal{U}_n} \mathbb{E}(\|D^r f - D^r f_0\|_\infty^2 | \mathbf{Y}, \sigma) \lesssim \prod_{k=1}^d J_k^{2r_k} \left(\frac{\log n}{n} \prod_{k=1}^d J_k + \sum_{k=1}^d J_k^{-2\alpha_k} \right).$$

To balance the orders of the two terms on the right, let $J_k = J^{1/\alpha_k}$ for $k = 1, \dots, d$. Then the bound above reduces to

$$O(J^{\sum_{k=1}^d (2r_k+1)/\alpha_k} n^{-1} \log n) + O(J^{2(\sum_{k=1}^d r_k/\alpha_k - 1)}) = O(\epsilon_{n,r,\infty}^2),$$

if $J \asymp (n/\log n)^{\alpha^*/(2\alpha^*+d)}$ and $J_k = J^{1/\alpha_k} \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \dots, d$. The rest of the proof can be completed as in Theorem 4.2. \square

PROOF OF THEOREM 5.1. Define $t_{n,r,\gamma_n}(\mathbf{x}) = \inf_{\sigma \in \mathcal{U}_n} z_{\gamma_n/2} \sigma \sqrt{\Sigma_r(\mathbf{x}, \mathbf{x})}$. To show $\hat{\mathcal{C}}_{n,r,\gamma_n}(\mathbf{x})$ has asymptotic coverage of 1, it suffices to show that

$$(7.23) \quad \sup_{\|f_0\|_{\alpha,\infty} \leq R} P_0 \left(|D^r f_0(\mathbf{x}) - D^r \tilde{f}(\mathbf{x})| > t_{n,r,\gamma_n}(\mathbf{x}) \right) \rightarrow 0.$$

Since $z_{\gamma_n/2} \rightarrow \infty$ and \mathcal{U}_n shrinks to σ_0 , we have $t_{n,r,\gamma_n}(\mathbf{x})^2 \gg \Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{x})$. In view of (3.11), $\Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{x})$ is bounded below by

$$\lambda_{\min}\{(\mathbf{B}^T \mathbf{B} + \mathbf{\Omega}^{-1})^{-1}\} \|\mathbf{W}_{\mathbf{r}}^T \mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{x})\|^2 \gtrsim n^{-1} \prod_{k=1}^d J_k \|\mathbf{W}_{\mathbf{r}}^T \mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{x})\|^2.$$

For any $\mathbf{x} = (x_1, \dots, x_d)^T \in [0, 1]^d$, let i_{x_k} be a positive integer such that $x_k \in [t_{k, i_{x_k}-1}, t_{k, i_{x_k}}]$. Then only $B_{i_{x_k}, q_k-r_k}(x_k), \dots, B_{i_{x_k}+q_k-r_k-1, q_k-r_k}(x_k)$ are nonzero at each $k = 1, \dots, d$. In view of (3.1), $(\partial^{r_k}/\partial x_k^{r_k})B_{j_k, q_k}(x_k)$ is a linear combination of $B_{j_k, q_k-r_k}(x_k), \dots, B_{j_k+r_k, q_k-r_k}(x_k)$ for any $1 \leq j_k \leq J_k$ with $k = 1, \dots, d$. Choose $j_k = i_{x_k} + q_k - r_k - 1$ for $k = 1, \dots, d$, then by (8.1), we have

$$\begin{aligned} \|\mathbf{W}_{\mathbf{r}}^T \mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{x})\|^2 &= \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \prod_{k=1}^d \left(\frac{\partial^{r_k}}{\partial x_k^{r_k}} B_{j_k, q_k}(x_k) \right)^2 \\ &\geq \prod_{k=1}^d \prod_{u=1}^{r_k} \frac{(q_k - u)^2 B_{i_{x_k}+q_k-r_k-1, q_k-r_k}(x_k)^2}{(t_{k, i_{x_k}+q_k-r_k-1} - t_{k, i_{x_k}-r_k-1+u})^2} \\ (7.24) \quad &\gtrsim \prod_{k=1}^d \frac{1}{\Delta_k^{2r_k}} \left(\frac{q_k - r_k}{q_k} \right)^2 \gtrsim \prod_{k=1}^d J_k^{2r_k}, \end{aligned}$$

since $t_{k, i_{x_k}+q_k-r_k-1} - t_{k, i_{x_k}-r_k-1+u} \leq (q_k - u + 1)\Delta_k \asymp J_k^{-1}$ for $k = 1, \dots, d$, by Lemma 8.1. Consequently, $t_{n,r,\gamma_n}(\mathbf{x})^2 \gg n^{-1} \prod_{k=1}^d J_k^{2r_k+1}$.

In view of (7.18), uniformly on $\|f_0\|_{\alpha, \infty} \leq R$,

$$(7.25) \quad \mathbb{E}_0[D^{\mathbf{r}} f_0(\mathbf{x}) - D^{\mathbf{r}} \tilde{f}(\mathbf{x})]^2 \lesssim n^{-1} \prod_{k=1}^d J_k^{2r_k+1} + \sum_{k=1}^d J_k^{-2(\alpha_k-r_k)}.$$

Hence uniformly on $\|f_0\|_{\alpha, \infty} \leq R$, the lack of coverage of $\hat{\mathcal{C}}_{n,r,\gamma_n}(\mathbf{x})$

$$P_0(|D^{\mathbf{r}} f_0(\mathbf{x}) - D^{\mathbf{r}} \tilde{f}(\mathbf{x})| > t_{n,r,\gamma_n}(\mathbf{x})) \lesssim \frac{n^{-1} \prod_{k=1}^d J_k^{2r_k+1} + \sum_{k=1}^d J_k^{-2(\alpha_k-r_k)}}{t_{n,r,\gamma_n}(\mathbf{x})^2}.$$

For the choice $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$, $k = 1, \dots, d$, the bound tends to zero uniformly on $\|f_0\|_{\alpha, \infty} \leq R$ and the diameter $\hat{\sigma}_n z_{\gamma_n/2} \sqrt{\Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{x})}$ of $\hat{\mathcal{C}}_{n,r,\gamma_n}(\mathbf{x})$ is $O_{P_0}(\epsilon_{n,r} \sqrt{\log(1/\gamma_n)})$ because $\Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{x}) \lesssim n^{-1} \prod_{k=1}^d J_k^{2r_k+1}$ by (7.16), $\hat{\sigma}_n$ converges to σ_0 and $z_{\gamma_n/2} = O(\sqrt{\log(1/\gamma_n)})$ by the estimate $P(Z > z) \leq z^{-1} \exp(-z^2/2)$ for $Z \sim N(0, 1)$.

To prove the corresponding assertion for the hierarchical Bayes credible interval, it suffices to show that

$$(7.26) \quad n^{-1} \prod_{k=1}^d J_k^{2r_k+1} \ll R_{n,r,\gamma_n}(\mathbf{x})^2 \ll n^{-1} \prod_{k=1}^d J_k^{2r_k+1} \log(1/\gamma_n)$$

uniformly on $\|f_0\|_{\alpha,\infty} \leq R$. If \mathcal{U}_n shrinks sufficiently slowly to σ_0 , we can ensure that with probability tending to one, $\Pi(\sigma \in \mathcal{U}_n | \mathbf{Y}) \geq 1 - \gamma_n$. By the definition of $R_{n,r,\gamma_n}(\mathbf{x})$, we have that

$$\begin{aligned} 1 - \gamma_n &= \Pi(|D^r f(\mathbf{x}) - A_r(\mathbf{x})\mathbf{Y} - c_r(\mathbf{x})\boldsymbol{\eta}| \leq R_{n,r,\gamma_n}(\mathbf{x}) | \mathbf{Y}) \\ &\leq \sup_{\sigma \in \mathcal{U}_n} \Pi(|D^r f(\mathbf{x}) - A_r(\mathbf{x})\mathbf{Y} - c_r(\mathbf{x})\boldsymbol{\eta}| \leq R_{n,r,\gamma_n}(\mathbf{x}) | \mathbf{Y}, \sigma) + \gamma_n. \end{aligned}$$

Since given σ , $z_{\gamma_n} \sigma \sqrt{\Sigma_r(\mathbf{x}, \mathbf{x})}$ is the $(1-2\gamma_n)$ -posterior quantile of $|D^r f(\mathbf{x}) - A_r(\mathbf{x})\mathbf{Y} - c_r(\mathbf{x})\boldsymbol{\eta}|$, it follows that on a set of probability tending to one, $R_{n,r,\gamma_n}(\mathbf{x}) \geq z_{\gamma_n} \inf\{\sigma : \sigma \in \mathcal{U}_n\} \sqrt{\Sigma_r(\mathbf{x}, \mathbf{x})}$. On the other hand from

$$1 - \gamma_n \geq \inf_{\sigma \in \mathcal{U}_n} \Pi(|D^r f(\mathbf{x}) - A_r(\mathbf{x})\mathbf{Y} - c_r(\mathbf{x})\boldsymbol{\eta}| \leq R_{n,r,\gamma_n}(\mathbf{x}) | \mathbf{Y}, \sigma) \Pi(\sigma \in \mathcal{U}_n | \mathbf{Y}),$$

we get $R_{n,r,\gamma_n}(\mathbf{x}) \leq z_{\gamma_n/2(1-\gamma_n)} \sup\{\sigma : \sigma \in \mathcal{U}_n\} \sqrt{\Sigma_r(\mathbf{x}, \mathbf{x})}$. This establishes (7.26). \square

PROOF OF THEOREM 5.3. For notational simplicity, we write $h_{n,r,\infty,\gamma}$ as $h_{\infty,\gamma}$ and define $t_{\infty,\gamma} = \inf_{\sigma \in \mathcal{U}_n} \sigma h_{n,r,\infty,\gamma}$. First we consider the empirical Bayes credible region. To show $\hat{\mathcal{C}}_{n,r,\infty,\gamma}^{\rho_n}$ has asymptotic coverage of 1, it suffices to show that

$$(7.27) \quad \sup_{\|f_0\|_{\alpha,\infty} \leq R} P_0(\|D^r f_0 - D^r \tilde{f}\|_{\infty} > \rho_n t_{\infty,\gamma}) \rightarrow 0.$$

Let $Z_{n,r} \sim \text{GP}(0, \Sigma_r)$. Let M_Z be the median of $\|Z_{n,r}\|_{\infty}$, i.e., M_Z satisfying $P(\|Z_{n,r}\|_{\infty} \leq M_Z) \geq 1/2$ and $P(\|Z_{n,r}\|_{\infty} \geq M_Z) \geq 1/2$. Let $\sigma_Z^2 = \sup_{\mathbf{x} \in [0,1]^d} \text{Var}(Z_{n,r}(\mathbf{x}))$ and note that by (7.16), $\sigma_Z^2 \lesssim n^{-1} \prod_{k=1}^d J_k^{2r_k+1} \rightarrow 0$ for $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha+d)\}}$, $k = 1, \dots, d$. Using the facts that $\sigma_Z \leq 2M_Z$ and $|\mathbb{E}\|Z_{n,r}\|_{\infty} - M_Z| \leq \sigma_Z(\pi/2)^{1/2}$ (see Pages 52 and 54 of [21]), we have $\mathbb{E}\|Z_{n,r}\|_{\infty} \asymp M_Z$.

Because $P(\|Z_{n,r}\|_{\infty} > h_{\infty,\gamma}) = \gamma$, and $\gamma < 1/2$, we have $h_{\infty,\gamma} \geq M_Z \asymp \mathbb{E}\|Z_{n,r}\|_{\infty}$. To lower bound $\mathbb{E}\|Z_{n,r}\|_{\infty}$, we introduce the notations $\mathcal{T}_k = \{t_{k,1}, \dots, t_{k,N_k}\}$, $k = 1, \dots, d$ and $\mathcal{T} = \prod_{k=1}^d \mathcal{T}_k$. Define $\mathcal{I} = \{(i_1, \dots, i_d) : 1 \leq i_k \leq N_k, k = 1, \dots, d\}$ and arrange the elements of \mathcal{I} lexicographically.

Then, we can enumerate the $N = \prod_{k=1}^d N_k$ elements of \mathcal{T} as $\{\boldsymbol{\tau}_i : i \in \mathcal{I}\}$, where $\boldsymbol{\tau}_i = (t_{1,i_1}, \dots, t_{d,i_d})$ with $(i_1, \dots, i_d) \in \mathcal{I}$. Define $u(x_1, \dots, x_d) = \prod_{k=1}^d (\partial^{r_k} / \partial x_k^{r_k}) B_{j_k, q_k}(x_k)$. Applying the multivariate mean value theorem to $u(x_1, \dots, x_d)$ at $\boldsymbol{\tau}_i$ and $\boldsymbol{\tau}_m$, we have for some point $\boldsymbol{\tau}^* = (t_1^*, \dots, t_d^*) = \lambda \boldsymbol{\tau}_i + (1 - \lambda) \boldsymbol{\tau}_m$ with $\lambda \in [0, 1]$,

$$\begin{aligned} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} |u(\boldsymbol{\tau}_i) - u(\boldsymbol{\tau}_m)|^2 &= \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} |\nabla u(\boldsymbol{\tau}^*)^T (\boldsymbol{\tau}_i - \boldsymbol{\tau}_m)|^2 \\ (7.28) \quad &= \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \left| \sum_{\beta=1}^d \left(\frac{\partial u}{\partial x_\beta} \right) (t_{\beta, i_\beta} - t_{\beta, m_\beta}) \right|^2. \end{aligned}$$

Choosing $j_1 = i_{x_1} + q_1 - r_1 - 2$ and $j_k = i_{x_k} + q_k - r_k - 1$ for $k = 2, \dots, d$, it then follows that $(\partial^{r_1+1} / \partial x_1^{r_1+1}) B_{j_1, q_1}(x_1) > 0$, while $(\partial^{r_k} / \partial x_k^{r_k}) B_{j_k, q_k}(x_k) > 0$ and $(\partial^{r_k+1} / \partial x_k^{r_k+1}) B_{j_k, q_k}(x_k) = 0$ for $k = 2, \dots, d$. We show only the first implication; the other two can be argued similarly. For $\mathbf{x} = (x_1, \dots, x_d)^T \in [0, 1]^d$, let i_{x_k} be a positive integer such that $x_k \in [t_{k, i_{x_k}-1}, t_{k, i_{x_k}}]$ for $k = 1, \dots, d$. Now by (3.1), $(\partial^{r_1+1} / \partial x_1^{r_1+1}) B_{j_1, q_1}(x_1)$ is a linear combination of the set of functions $\{B_{j_1, q_1-r_1-1}(x_1), \dots, B_{j_1+r_1+1, q_1-r_1-1}(x_1)\}$ while only $\{B_{i_{x_1}, q_1-r_1-1}(x_1), \dots, B_{i_{x_1}+q_1-r_1-2, q_1-r_1-1}(x_1)\}$ are nonzero by the support property of B-splines. For $j_1 = i_{x_1} + q_1 - r_1 - 2$, only the positive term corresponding to $B_{i_{x_1}+q_1-r_1-2, q_1-r_1-1}(x_1)$, with coefficients given by the second equation of (8.1) below survives. Thus, only $\partial u / \partial x_1$ will be positive while $\partial u / \partial x_k = 0$ for $k = 2, \dots, d$. By repeated applications of (7.24), the right hand side of (7.28) is bounded below by

$$\begin{aligned} &\left(\frac{\partial^{r_1+1}}{\partial x_1^{r_1+1}} B_{j_1, q_1}(x_1) \right)^2 \prod_{k=2}^d \left(\frac{\partial^{r_k}}{\partial x_k^{r_k}} B_{j_k, q_k}(x_k) \right)^2 (t_{1, i_1} - t_{1, m_1})^2 \\ &\gtrsim J_1^{2r_1+2} \prod_{k=2}^d J_k^{2r_k} \left(\min_{1 \leq l \leq N_1} \delta_{1, l}^2 \right) \gtrsim \prod_{k=1}^d J_k^{2r_k}, \end{aligned}$$

where $\delta_{1, l} = t_{1, l} - t_{1, l-1}$ for $1 \leq l \leq N_1$, and the last inequality follows from the quasi-uniformity of knots and Lemma 8.1. Define $V_i = Z_{n, \mathbf{r}}(\boldsymbol{\tau}_i)$ for $\mathbf{1}_d \leq \mathbf{i} \leq \mathbf{N}$ where $\mathbf{N} = (N_1, \dots, N_d)^T$. Note that $\|Z_{n, \mathbf{r}}\|_\infty \geq \max_{\mathbf{1}_d \leq \mathbf{i} \leq \mathbf{N}} V_i$. Then by (3.11), for any $\mathbf{1}_d \leq \mathbf{i}, \mathbf{m} \leq \mathbf{N}$,

$$\begin{aligned} E(V_i - V_m)^2 &\geq \lambda_{\min}\{(\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1}\} \|\mathbf{W}_{\mathbf{r}}^T(\mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\boldsymbol{\tau}_i) - \mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\boldsymbol{\tau}_m))\|^2 \\ &\gtrsim \left(\frac{1}{n} \prod_{k=1}^d J_k \right) \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} |u(\boldsymbol{\tau}_i) - u(\boldsymbol{\tau}_m)|^2 \geq \frac{c}{n} \prod_{k=1}^d J_k^{2r_k+1}, \end{aligned}$$

for a universal constant $c > 0$. Define $U_{\mathbf{i}} = \sqrt{(2n/c)} \prod_{k=1}^d J_k^{-(r_k+1/2)} V_{\mathbf{i}}$ and let $H_{\mathbf{i}}$ be i.i.d. $N(0, 1)$ with $\mathbf{1}_d \leq \mathbf{i} \leq \mathbf{N}$. By (3.14) of [21], we have $E(\max_{\mathbf{1}_d \leq \mathbf{i} \leq \mathbf{N}} H_{\mathbf{i}}) \gtrsim \sqrt{\log N}$. Now, $E(U_{\mathbf{i}} - U_{\mathbf{m}})^2 \geq 2 = E(H_{\mathbf{i}} - H_{\mathbf{m}})^2$ and hence by Slepian's Lemma (Corollary 3.14 of [21]),

$$E\left(\max_{\mathbf{1}_d \leq \mathbf{i} \leq \mathbf{N}} U_{\mathbf{i}}\right) \geq E\left(\max_{\mathbf{1}_d \leq \mathbf{i} \leq \mathbf{N}} H_{\mathbf{i}}\right) \gtrsim \sqrt{\log N},$$

where $N = \prod_{k=1}^d N_k \sim \prod_{k=1}^d J_k$. It then follows that $t_{\infty, \gamma}^2 \gtrsim \sigma_0^2 h_{\infty, \gamma}^2 \gtrsim (\log n/n) \prod_{k=1}^d J_k^{2r_k+1}$. Therefore using (7.21) and (7.22), we have uniformly on $\|f_0\|_{\alpha, \infty} \leq R$,

$$(7.29) \quad \begin{aligned} E_0(\|D^{\mathbf{r}} f_0 - D^{\mathbf{r}} \tilde{f}\|_{\infty}^2) &\leq 2E(\|Q_{n, \mathbf{r}}\|_{\infty}^2) + 2\|\mathbf{A}_{\mathbf{r}} \mathbf{F}_0 + \mathbf{c}_{\mathbf{r}} \boldsymbol{\eta} - D^{\mathbf{r}} f_0\|_{\infty}^2 \\ &\lesssim \frac{\log n}{n} \prod_{k=1}^d J_k^{2r_k+1} + \sum_{k=1}^d J_k^{-2(\alpha_k - r_k)}. \end{aligned}$$

Hence for the choice $J_k \asymp (n/\log n)^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$, $k = 1, \dots, d$, $P_0(\|D^{\mathbf{r}} f_0 - D^{\mathbf{r}} \tilde{f}\|_{\infty} > \rho_n t_{\infty, \gamma}) \rightarrow 0$ since $t_{\infty, \gamma}^2 \gtrsim (\log n/n) \prod_{k=1}^d J_k^{2r_k+1}$ and $\rho_n \rightarrow \infty$.

If the true errors are i.i.d. $N(0, \sigma_0^2)$, then $Q_{n, \mathbf{r}} \sim \text{GP}(0, \sigma_0^2 \Psi_{\mathbf{r}})$ under P_0 . Define $\sigma_Q^2 = \sup_{\mathbf{x} \in [0, 1]^d} \text{Var}(Q_{n, \mathbf{r}}(\mathbf{x}))$. We have for constants $C_1, C_2, C_3 > 0$,

$$t_{\infty, \gamma} \geq C_1 \epsilon_{n, \mathbf{r}, \infty}, \quad \|D^{\mathbf{r}} f_0 - E_0 D^{\mathbf{r}} \tilde{f}\|_{\infty} \leq C_2 \epsilon_{n, \mathbf{r}, \infty}, \quad E\|Q_{n, \mathbf{r}}\|_{\infty} \leq C_3 \epsilon_{n, \mathbf{r}, \infty}.$$

The first inequality was established above, while the second and third inequalities follow from (7.22) and (7.21). Then by Proposition A.2.1 of [35], $P_0(\|Q_{n, \mathbf{r}}\|_{\infty} > 2C_3 \epsilon_{n, \mathbf{r}, \infty})$ is bounded by

$$P_0(\|Q_{n, \mathbf{r}}\|_{\infty} > E\|Q_{n, \mathbf{r}}\|_{\infty} + C_3 \epsilon_{n, \mathbf{r}, \infty}) \leq 2 \exp\{-C_3^2 \epsilon_{n, \mathbf{r}, \infty}^2 / (2\sigma_Q^2)\}.$$

In view of (7.17), we have $\sigma_Q^2 = O(\epsilon_{n, \mathbf{r}}^2)$. Since $\epsilon_{n, \mathbf{r}} \ll \epsilon_{n, \mathbf{r}, \infty}$, this implies that the right hand side above tends to zero as $n \rightarrow \infty$. By the triangle inequality, we have

$$P_0(\|D^{\mathbf{r}} f_0 - D^{\mathbf{r}} \tilde{f}\|_{\infty} > \rho t_{\infty, \gamma}) \leq P_0(\|Q_{n, \mathbf{r}}\|_{\infty} > \rho t_{\infty, \gamma} - \|D^{\mathbf{r}} f_0 - E_0 D^{\mathbf{r}} \tilde{f}\|_{\infty})$$

which tends to 0 if $\rho \geq (2C_3 + C_2)/C_1$.

To estimate the diameter $\hat{\sigma}_n \rho_n h_{\infty, \gamma}$, the last inequality in Proposition A.2.1 of [35] gives $\gamma = P(\|Z_{n, \mathbf{r}}\|_{\infty} > h_{\infty, \gamma}) \leq 2 \exp\{-h_{\infty, \gamma}^2 / (8E\|Z_{n, \mathbf{r}}\|_{\infty}^2)\}$. Therefore, $h_{\infty, \gamma} \lesssim (E\|Z_{n, \mathbf{r}}\|_{\infty}^2)^{1/2} \sqrt{-\log \gamma}$ and hence the assertion follows from (7.21).

To prove the assertions about hierarchical Bayes credible regions, we proceed as in the proof of Theorem 5.1. By definition

$$\begin{aligned} 1 - \gamma &= \Pi(\|D^{\mathbf{r}} f - A_{\mathbf{r}} \mathbf{Y} - c_{\mathbf{r}} \boldsymbol{\eta}\|_{\infty} \leq R_{n,\mathbf{r},\infty,\gamma} | \mathbf{Y}) \\ &\leq \sup_{\sigma \in \mathcal{U}_n} \Pi(\|D^{\mathbf{r}} f - A_{\mathbf{r}} \mathbf{Y} - c_{\mathbf{r}} \boldsymbol{\eta}\|_{\infty} \leq R_{n,\mathbf{r},\infty,\gamma} | \mathbf{Y}, \sigma) + \Pi(\sigma \notin \mathcal{U}_n | \mathbf{Y}). \end{aligned}$$

Choose γ' strictly between γ and $1/2$. Making \mathcal{U}_n to shrink sufficiently slowly to σ_0 so that $\Pi(\sigma \in \mathcal{U}_n | \mathbf{Y}) \geq 1 - \gamma' + \gamma$ with probability tending to one and using the facts that the conditional posterior distribution of $(D^{\mathbf{r}} f - A_{\mathbf{r}} \mathbf{Y} - c_{\mathbf{r}} \boldsymbol{\eta})/\sigma$ given σ is equal to the distribution of the Gaussian process $Z_{n,\mathbf{r}}$, which is free of σ , and $\|Z_{n,\mathbf{r}}\|_{\infty}$ has $(1 - \gamma')$ quantile $t_{\infty,\gamma'}$, we obtain

$$R_{n,\mathbf{r},\infty,\gamma} \geq \inf\{\sigma : \sigma \in \mathcal{U}_n\} t_{\infty,\gamma'} \asymp t_{\infty,\gamma'} \gtrsim \epsilon_{n,\mathbf{r},\infty}.$$

Hence the modified hierarchical Bayes credible region $\mathcal{C}_{n,\mathbf{r},\infty,\gamma}^{\rho_n}$ has asymptotic coverage 1 for any $\rho_n \rightarrow \infty$, and for the Gaussian true error we can choose $\rho_n = \rho$ for a sufficiently large constant. To bound the diameter of $\mathcal{C}_{n,\mathbf{r},\infty,\gamma}^{\rho_n}$ we use the relation

$$1 - \gamma \geq \inf_{\sigma \in \mathcal{U}_n} \Pi(\|D^{\mathbf{r}} f - A_{\mathbf{r}} \mathbf{Y} - c_{\mathbf{r}} \boldsymbol{\eta}\|_{\infty} \leq R_{n,\mathbf{r},\infty,\gamma} | \mathbf{Y}, \sigma) \Pi(\sigma \in \mathcal{U}_n | \mathbf{Y})$$

to conclude that $R_{n,\mathbf{r},\infty,\gamma} \leq t_{\infty,\gamma/(1-\gamma'+\gamma)}$, which is of the order $\epsilon_{n,\mathbf{r},\infty}$ since $\gamma/(1 - \gamma' + \gamma) < 1/2$ by the choice of γ' . \square

PROOF OF REMARK 5.2. We indicate how to show coverage of the empirical Bayes credible region; the necessary changes for the hierarchical version can be made as in the proofs of Theorems 5.1 and 5.3. The adequacy of the coverage will be shown if $P_0\left(\|D^{\mathbf{r}} f_0 - D^{\mathbf{r}} \tilde{f}\|_2 > t_{n,\mathbf{r},2,\gamma_n}\right) \rightarrow 0$ uniformly on $\|f_0\|_{\alpha,\infty} \leq R$, where $t_{n,\mathbf{r},2,\gamma_n} = \inf_{\sigma^2 \in \mathcal{U}_n} \sigma h_{n,\mathbf{r},2,\gamma_n}$. Let $Z_{n,\mathbf{r}} \sim \text{GP}(0, \Sigma_{\mathbf{r}})$. Since $\|Z_{n,\mathbf{r}}\|_2 \geq \int Z_{n,\mathbf{r}}$ which is normally distributed with mean 0 and variance $\int \int \Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y}$, it follows from (3.6) that

$$h_{n,\mathbf{r},2,\gamma_n}^2 \gg \int \int \Sigma_{\mathbf{r}}(\mathbf{x}, \mathbf{y}) d\mathbf{x} d\mathbf{y} \gtrsim \frac{1}{n} \prod_{k=1}^d J_k \left\| \int \mathbf{W}_{\mathbf{r}}^T \mathbf{b}_{J,\mathbf{q}-\mathbf{r}}(\mathbf{x}) d\mathbf{x} \right\|^2.$$

Extending the last two equations in the proof of Lemma 6.7 in [36] to multivariate splines by arguments used in the proof of the last two theorems, it follows from the last display that $t_{n,\mathbf{r},2,\gamma_n}^2 \gtrsim \sigma_0^2 h_{n,\mathbf{r},2,\gamma_n}^2 \gg n^{-1} \prod_{k=1}^d J_k^{2r_k+1}$. On the other hand,

$$\mathbb{E}_0 \|f_0 - \tilde{f}\|_2^2 = \int \mathbb{E}_0 |f_0(\mathbf{x}) - \tilde{f}(\mathbf{x})|^2 d\mathbf{x} \lesssim \frac{1}{n} \prod_{k=1}^d J_k^{2r_k+1} + \sum_{k=1}^d J_k^{-2(\alpha_k - r_k)},$$

by (7.25). Then uniformly on $\|f_0\|_{\alpha,\infty} \leq R$, the coverage of $\hat{C}_{n,r,2,\gamma_n}$ goes to one in probability by Markov's inequality. Since in view of (3.5) from [21], $\gamma_n = P(\|Z_{n,r}\|_2 > h_{n,r,2,\gamma_n}) \leq 4 \exp\{-h_{n,r,2,\gamma_n}^2 / (8E\|Z_{n,r}\|_2^2)\}$, the size of the radius of the L_2 -confidence region is estimated as $O_{P_0}(\epsilon_{n,r} \sqrt{\log(1/\gamma_n)})$. \square

8. Appendix.

LEMMA 8.1. *Under quasi-uniform knots, $\Delta_k \asymp N_k^{-1} \asymp J_k^{-1}$, $k = 1, \dots, d$.*

PROOF. The proof is straightforward because all N_k spacings are of the same order and they sum to one. \square

LEMMA 8.2. *Each non-zero entry of \mathbf{W}_r defined implicitly in (3.2) is uniformly $O(\prod_{k=1}^d \Delta_k^{-r_k})$.*

PROOF. Recall that the dimension of \mathbf{W}_r is $\prod_{k=1}^d (J_k - r_k) \times \prod_{k=1}^d J_k$. In view of (3.1), each row of \mathbf{W}_r has only $\prod_{k=1}^d (r_k + 1)$ nonzero entries and their arrangement is analogous to a banded matrix, namely the position of nonzero entries in the current row is a shift of one entry to the right of the nonzero entries' position in the previous row. Also, each column of \mathbf{W}_r has at most $\prod_{k=1}^d (r_k + 1)$ nonzero entries. We index the rows and columns of \mathbf{W}_r using d -dimensional indices as in Definition 2.2.

Define $\tilde{r} = \prod_{k=1}^d (r_k + 1) - 2$, and let $\mathcal{G} = \{\mathbf{u} = (u_1, \dots, u_d) : \mathbf{0} \leq \mathbf{u} \leq \mathbf{r}, \mathbf{u} \neq \mathbf{r}, \mathbf{u} \neq \mathbf{0}\}$. By ordering the elements in \mathcal{G} lexicographically, we can enumerate its elements by $\mathcal{G} = \{\mathbf{g}_1, \dots, \mathbf{g}_{\tilde{r}}\}$. Furthermore, define sets $\mathcal{I} = \{(i_1, \dots, i_d) : 1 \leq i_k \leq J_k - r_k, k = 1, \dots, d\}$ and $\mathcal{J} = \{(j_1, \dots, j_d) : 1 \leq j_k \leq J_k, k = 1, \dots, d\}$, where we order their elements lexicographically. Let $w_{i,j}^{(r)}$ denote the (i, j) th element of \mathbf{W}_r such that $\mathbf{i} \in \mathcal{I}$ and $\mathbf{j} \in \mathcal{J}$. The expressions for the nonzero entries can be described as follows: for each row $\mathbf{i} \in \mathcal{I}$, the first and last nonzero entries are given by

$$(8.1) \quad \begin{aligned} w_{i,i}^{(r)} &= (-1)^{\sum_{k=1}^d r_k} \prod_{k=1}^d \prod_{l=1}^{r_k} \frac{q_k - l}{t_{k,i_k} - t_{k,i_k - q_k + l}}, \\ w_{i,i+r}^{(r)} &= \prod_{k=1}^d \prod_{l=1}^{r_k} \frac{q_k - l}{t_{k,i_k + l - 1} - t_{k,i_k - q_k + r_k}}. \end{aligned}$$

If \tilde{r} is odd, we partition $\mathcal{G} = \mathcal{G}_1 \cup \{\mathbf{g}_{(\tilde{r}+1)/2}\} \cup \mathcal{G}_2$ where $\mathcal{G}_1 = \{\mathbf{g}_1, \dots, \mathbf{g}_{(\tilde{r}-1)/2}\}$ and $\mathcal{G}_2 = \{\mathbf{g}_{(\tilde{r}+3)/2}, \dots, \mathbf{g}_{\tilde{r}}\}$. The intermediate nonzero entries $w_{i,i+h}^{(r)}$ for

$\mathbf{h} = (h_1, \dots, h_d)^T \in \mathcal{G}_1$ are

$$(8.2) \quad (-1)^{\sum_{k=1}^d (r_k - h_k)} w_{\mathbf{i}, \mathbf{i}}^{(r)} \left[1 + \prod_{k=1}^d \sum_{t=1}^{\binom{r_k}{h_k} - 1} \prod_{s=1}^t \frac{t_{k, i_k} - t_{k, i_k - q_k + s}}{t_{k, i_k + 1} - t_{k, i_k + 1 - q_k + s}} \right],$$

while for $\mathbf{h} \in \mathcal{G}_2$, it is

$$(8.3) \quad (-1)^{\sum_{k=1}^d (r_k - h_k)} w_{\mathbf{i}, \mathbf{i} + \mathbf{r}}^{(r)} \left[1 + \prod_{k=1}^d \sum_{t=1}^{\binom{r_k}{h_k} - 1} \prod_{s=1}^t \frac{t_{k, i_k + r_k - s} - t_{k, i_k + r_k - q_k}}{t_{k, i_k + r_k - 1 - s} - t_{k, i_k + r_k - 1 - q_k}} \right],$$

When $\mathbf{h} = \mathbf{g}_{(\tilde{r}+1)/2}$, we have

$$(8.4) \quad \begin{aligned} w_{\mathbf{i}, \mathbf{i} + \mathbf{h}}^{(r)} &= w_{\mathbf{i}, \mathbf{i}}^{(r)} \prod_{k=1}^d \left(\frac{t_{k, i_k} - t_{k, i_k - q_k + 1}}{t_{k, i_k + 1} - t_{k, i_k + 2 - q_k}} \right) \\ &\quad \times \left[1 + \prod_{k=1}^d \sum_{t=1}^{\binom{r_k - 1}{r_k/2 - 1} - 1} \prod_{s=0}^{t-1} \frac{t_{k, i_k + s} - t_{k, i_k + 2 - q_k}}{t_{k, i_k + 1 + s} - t_{k, i_k + 3 - q_k}} \right] \\ &\quad + w_{\mathbf{i}, \mathbf{i} + \mathbf{r}}^{(r)} \prod_{k=1}^d \left(\frac{t_{k, i_k + r_k - 1} - t_{k, i_k + r_k - q_k}}{t_{k, i_k + r_k - 2} - t_{k, i_k + r_k - 1 - q_k}} \right) \\ &\quad \times \left[1 + \prod_{k=1}^d \sum_{t=1}^{\binom{r_k - 1}{r_k/2} - 1} \prod_{s=0}^{t-1} \frac{t_{k, i_k + r_k - 2} - t_{k, i_k + r_k - q_k - s}}{t_{k, i_k + r_k - 3} - t_{k, i_k + r_k - 1 - q_k - s}} \right]. \end{aligned}$$

If \tilde{r} is even, we partition $\mathcal{G} = \mathcal{G}_1 \cup \mathcal{G}_2$ where $\mathcal{G}_1 = \{\mathbf{g}_1, \dots, \mathbf{g}_{\tilde{r}/2}\}$ and $\mathcal{G}_2 = \{\mathbf{g}_{\tilde{r}/2+1}, \dots, \mathbf{g}_{\tilde{r}}\}$. Then the expression for $w_{\mathbf{i}, \mathbf{i} + \mathbf{h}}^{(r)}$ is (8.2) for $\mathbf{h} \in \mathcal{G}_1$ and is (8.3) for $\mathbf{h} \in \mathcal{G}_2$. By the quasi-uniformity of the knots, the endpoints in (8.1) are $O(\prod_{k=1}^d \Delta_k^{-r_k})$, while the fractions of knot differences appearing in (8.2)–(8.4) are $O(1)$. \square

LEMMA 8.3. $\sum_{i=1}^n \prod_{k=1}^d B_{j_k, q_k}(X_{ik})^{p_k} \lesssim n \prod_{k=1}^d J_k^{-1}$ for $1 \leq j_k \leq J_k$ and $p_k \in \mathbb{N}$, $k = 1, \dots, d$.

PROOF. As $B_{j_k, q_k}(\cdot) \leq 1$ and is positive only inside $(t_{k, j_k - q_k}, t_{k, j_k})$,

$$\sum_{i=1}^n \prod_{k=1}^d B_{j_k, q_k}(X_{ik})^{p_k} \leq n \int_{[0,1]^d} \prod_{k=1}^d \mathbb{1}_{(t_{k, j_k - q_k}, t_{k, j_k}]}(\mathbf{x}) dG_n(\mathbf{x}).$$

By the quasi-uniformity of the knots, we have $t_{k,j_k} - t_{k,j_k-q_k} \leq q_k \Delta_k$ and $t_{k,j_k} - t_{k,j_k-q_k} \geq q_k \min_{1 \leq l \leq N_k} \delta_{k,l} \geq q_k \Delta_k / C$. This implies that $t_{k,j_k} - t_{k,j_k-q_k} \asymp \Delta_k$ for $k = 1, \dots, d$. Moreover, assumption (2.2) and Lemma 8.1 imply that the right hand side above is

$$\begin{aligned} nG_n \left[\prod_{k=1}^d (t_{k,j_k-q_k}, t_{k,j_k}) \right] &= nG \left[\prod_{k=1}^d (t_{k,j_k-q_k}, t_{k,j_k}) \right] + o \left(n \prod_{k=1}^d N_k^{-1} \right) \\ &\lesssim n \prod_{k=1}^d \Delta_k + o \left(n \prod_{k=1}^d \Delta_k \right) \lesssim n \prod_{k=1}^d J_k^{-1}. \quad \square \end{aligned}$$

LEMMA 8.4. *Let \mathbf{A} be a $J \times J$ symmetric and positive definite matrix with its rows and columns indexed by d -dimensional multi-indices, i.e., for $\mathbf{i} = (i_1, \dots, i_d)$ and $\mathbf{j} = (j_1, \dots, j_d)$, such that $1 \leq i_k, j_k \leq J_k, k = 1, \dots, d, J = \prod_{k=1}^d J_k$, the (\mathbf{i}, \mathbf{j}) th element of \mathbf{A} is $a_{\mathbf{i}, \mathbf{j}} = \mathbf{A}\{(\mathbf{i}_1, \dots, \mathbf{i}_d), (\mathbf{j}_1, \dots, \mathbf{j}_d)\}$. Let \mathbf{A} be $\mathbf{q} = (q_1, \dots, q_d)^T$ banded as in Definition 2.2. Furthermore, assume that the eigenvalues of \mathbf{A} are contained in $[a\tau_m, b\tau_m]$ for fixed $0 < a < b < \infty$ and some sequence τ_m . Then $\|\mathbf{A}^{-1}\|_{(\infty, \infty)} = O(\tau_m^{-1})$.*

PROOF. We adapt the proof given in Proposition 2.2 of [11] to the case of multi-dimensional banded matrix. We first note that if \mathbf{A} is \mathbf{q} -banded and \mathbf{B} is \mathbf{w} -banded as in Definition 2.2, then \mathbf{AB} is $\mathbf{q} + \mathbf{w}$ banded. To see this, observe that $(\mathbf{AB})_{\mathbf{i}, \mathbf{j}} = \sum_{l_1=1}^{J_1} \dots \sum_{l_d=1}^{J_d} a_{(\mathbf{i}_1, \dots, \mathbf{i}_d), (l_1, \dots, l_d)} b_{(\mathbf{j}_1, \dots, \mathbf{j}_d), (l_1, \dots, l_d)} \neq 0$ only if at least one of the terms in the sum is nonzero. Thus $a_{(\mathbf{i}_1, \dots, \mathbf{i}_d), (l_1, \dots, l_d)} \neq 0$ and $b_{(\mathbf{j}_1, \dots, \mathbf{j}_d), (l_1, \dots, l_d)} \neq 0$ for some (l_1, \dots, l_d) . Hence $|i_k - l_k| \leq q_k$ and $|j_k - l_k| \leq w_k$ for $k = 1, \dots, d$, and by the triangle inequality, $|i_k - j_k| \leq q_k + w_k$ for $k = 1, \dots, d$. Therefore, \mathbf{AB} is $\mathbf{q} + \mathbf{w}$ banded. Repeated applications of the same argument show that \mathbf{A}^n is $n\mathbf{q}$ -banded.

Since we can scale \mathbf{A} by τ_m such that its eigenvalues are in $[a, b]$, we set $\tau_m = 1$ without loss of generality. Let $p_n(\cdot)$ be a polynomial of degree n . Then $p_n(\mathbf{A})$ is $n\mathbf{q}$ -banded. Since the set of eigenvalues for \mathbf{A} is $\Lambda(\mathbf{A}) \subseteq [a, b]$ by assumption, spectral theorem and Proposition 2.1 of [11] imply that

$$\|\mathbf{A}^{-1} - p_n(\mathbf{A})\|_{(2,2)} = \max_{x \in \Lambda(\mathbf{A})} |1/x - p_n(x)| \leq C_0 [(\sqrt{b/a} - 1)/(\sqrt{b/a} + 1)]^{n+1}$$

for $C_0 = (1 + \sqrt{b/a})^2/(2b)$. For any $n \in \mathbb{N}$, $p_n(\mathbf{A})_{\mathbf{i}, \mathbf{j}} = 0$ if $|i_k - j_k| > nq_k$ for some $1 \leq k \leq d$. Suppose $\mathbf{i} \neq \mathbf{j}$, choose n to satisfy $n < \max_{1 \leq k \leq d} |i_k - j_k|$

$j_k |q_k^{-1} \leq n + 1$. Therefore,

$$\begin{aligned}
 |\mathbf{A}^{-1}(\mathbf{i}, \mathbf{j})| &= |\mathbf{A}^{-1}(\mathbf{i}, \mathbf{j}) - p_n(\mathbf{A})_{\mathbf{i}, \mathbf{j}}| \leq \|\mathbf{A}^{-1} - p_n(\mathbf{A})\|_{(2,2)} \\
 &\leq C_0 [(\sqrt{b/a} - 1)/(\sqrt{b/a} + 1)]^{\max_{1 \leq k \leq d} |i_k - j_k| / q_k} \\
 (8.5) \quad &\leq C_0 \lambda^{\sum_{k=1}^d |i_k - j_k|},
 \end{aligned}$$

where $\lambda = [(\sqrt{b/a} - 1)/(\sqrt{b/a} + 1)]^{1/\sum_{k=1}^d q_k}$. When $i_k = j_k$ for all $k = 1, \dots, d$, we have $\mathbf{A}^{-1}(\mathbf{i}, \mathbf{i}) \leq \|\mathbf{A}^{-1}\|_{(2,2)} = 1/\lambda_{\min}(\mathbf{A}) \leq 1/a$. Combining this case with (8.5), we have $|\mathbf{A}^{-1}(\mathbf{i}, \mathbf{j})| \leq C \lambda^{\sum_{k=1}^d |i_k - j_k|}$ for $C = \max\{C_0, 1/a\}$. Since $0 < \lambda < 1$,

$$\begin{aligned}
 \|\mathbf{A}^{-1}\|_{(\infty, \infty)} &\leq C \max_{1 \leq i_k \leq J_k, k=1, \dots, d} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \prod_{k=1}^d \lambda^{|i_k - j_k|} \\
 &\lesssim \prod_{k=1}^d \left(1 + 2 \sum_{j_k=1}^{J_k} \lambda^{j_k} \right) \lesssim \left(1 + 2 \sum_{j=1}^{\infty} \lambda^j \right)^d < \infty. \quad \square
 \end{aligned}$$

LEMMA 8.5. $\|\mathbf{b}_{J, q-r}(\mathbf{x}) - \mathbf{b}_{J, q-r}(\mathbf{y})\|^2 \lesssim \|\mathbf{J}\|^2 \|\mathbf{x} - \mathbf{y}\|^2$ for $\mathbf{x}, \mathbf{y} \in [0, 1]^d$.

PROOF. By equation (8) of Chapter X in [8] and the triangle inequality,

$$\begin{aligned}
 |B'_{j_k, q_k - r_k}(x_k)| &\lesssim \frac{|B_{j_k, q_k - r_k - 1}(x_k)|}{t_{k, j_k + q_k - r_k - 1} - t_{k, j_k}} + \frac{|B_{j_k + 1, q_k - r_k - 1}(x_k)|}{t_{k, j_k + q_k - r_k} - t_{k, j_k + 1}} \\
 (8.6) \quad &\lesssim \left(\min_{1 \leq l \leq N_k} \delta_{k, l} \right)^{-1} \lesssim \Delta_k^{-1} \lesssim J_k,
 \end{aligned}$$

where we have used the quasi-uniformity of the knots and Lemma 8.1. Using $|\prod_{i=1}^d a_i - \prod_{i=1}^d b_i| \leq \sum_{i=1}^d |a_i - b_i|$ for $|a_i| \leq 1, |b_i| \leq 1, i = 1, \dots, d$, the mean value theorem, (8.6) and the Cauchy-Schwarz inequality,

$$\begin{aligned}
 \left| \prod_{k=1}^d B_{j_k, q_k - r_k}(x_k) - \prod_{k=1}^d B_{j_k, q_k - r_k}(y_k) \right| &\leq \sum_{k=1}^d |B_{j_k, q_k - r_k}(x_k) - B_{j_k, q_k - r_k}(y_k)| \\
 &\lesssim \sum_{k=1}^d J_k |x_k - y_k| \leq \|\mathbf{J}\| \|\mathbf{x} - \mathbf{y}\|.
 \end{aligned}$$

Since at most $2 \prod_{k=1}^d (q_k - r_k)$ elements in both $\mathbf{b}_{J, q-r}(\mathbf{x})$ and $\mathbf{b}_{J, q-r}(\mathbf{y})$

will be nonzero for any $\mathbf{x}, \mathbf{y} \in [0, 1]^d$, $\|\mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{x}) - \mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{y})\|^2$ is

$$\begin{aligned} & \sum_{j_1=1}^{J_1-r_1} \cdots \sum_{j_d=1}^{J_d-r_d} \left| \prod_{k=1}^d B_{j_k, q_k-r_k}(x_k) - \prod_{k=1}^d B_{j_k, q_k-r_k}(y_k) \right|^2 \\ & \lesssim \left[2 \prod_{k=1}^d (q_k - r_k) \right] \sum_{k=1}^d J_k^2 \|\mathbf{x} - \mathbf{y}\|^2 \lesssim \sum_{k=1}^d J_k^2 \|\mathbf{x} - \mathbf{y}\|^2. \quad \square \end{aligned}$$

LEMMA 8.6. *Let $\mathbf{r} \in \mathbb{N}_0^d$ be such that $\sum_{k=1}^d r_k/\alpha_k < 1$. Let $Z_{n,\mathbf{r}} \sim \text{GP}(0, \Sigma_{\mathbf{r}})$ and $Q_{n,\mathbf{r}}$ be a sub-Gaussian process with mean function 0 and covariance function $\sigma_0^2 \Psi_{\mathbf{r}}$. Let $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \dots, d$. Then for any $\mathbf{t}, \mathbf{s} \in [0, 1]^d$, we have $\text{Var}[Z_{n,\mathbf{r}}(\mathbf{t}) - Z_{n,\mathbf{r}}(\mathbf{s})] \leq C\|\mathbf{J}\|^2\|\mathbf{t} - \mathbf{s}\|^2$ and $\text{Var}[Q_{n,\mathbf{r}}(\mathbf{t}) - Q_{n,\mathbf{r}}(\mathbf{s})] \leq C\|\mathbf{J}\|^2\|\mathbf{t} - \mathbf{s}\|^2$ for some constant $C > 0$.*

PROOF. Let $J_k \asymp n^{\alpha^*/\{\alpha_k(2\alpha^*+d)\}}$ for $k = 1, \dots, d$, then $\text{Var}[Z_{n,\mathbf{r}}(\mathbf{t}) - Z_{n,\mathbf{r}}(\mathbf{s})]$ is bounded above by

$$\begin{aligned} & \|\mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{t}) - \mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{s})\|^2 \left\| (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1} \right\|_{(2,2)}^2 \|\mathbf{W}_{\mathbf{r}}^T \mathbf{W}_{\mathbf{r}}\|_{(2,2)} \\ & \lesssim \frac{1}{n} \left(\prod_{k=1}^d J_k^{2r_k+1} \right) \left(\sum_{k=1}^d J_k^2 \right) \|\mathbf{t} - \mathbf{s}\|^2 \lesssim \|\mathbf{J}\|^2 \|\mathbf{t} - \mathbf{s}\|^2, \end{aligned}$$

where we used Lemma 8.5, equations (3.11) and (7.15) to bound the three norms respectively. Similarly, $\text{Var}[Q_{n,\mathbf{r}}(\mathbf{t}) - Q_{n,\mathbf{r}}(\mathbf{s})]$ is bounded by

$$\|\mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{t}) - \mathbf{b}_{\mathbf{J}, \mathbf{q}-\mathbf{r}}(\mathbf{s})\|^2 \left\| (\mathbf{B}^T \mathbf{B} + \boldsymbol{\Omega}^{-1})^{-1} \right\|_{(2,2)}^2 \|\mathbf{B}^T \mathbf{B}\|_{(2,2)} \|\mathbf{W}_{\mathbf{r}}^T \mathbf{W}_{\mathbf{r}}\|_{(2,2)},$$

which is $O(\|\mathbf{J}\|^2\|\mathbf{t} - \mathbf{s}\|^2)$, where we used Lemma 8.5, (3.11), (3.10) and (7.15) to bound the four norms respectively. \square

LEMMA 8.7. *Let $f(\mathbf{x}) = \mathbf{b}_{\mathbf{J}, \mathbf{q}}(\mathbf{x})^T \boldsymbol{\theta}$ and $I_{j_1, \dots, j_d} = \prod_{k=1}^d [t_{k, j_k - q_k}, t_{k, j_k}]$. Furthermore, let $f|_{I_{j_1, \dots, j_d}}$ be the restriction of f onto I_{j_1, \dots, j_d} . Then there exists constant $C > 0$ depending on $\mathbf{q} = (q_1, \dots, q_d)^T$ such that*

$$\|f|_{I_{j_1, \dots, j_d}}\|_{\infty} \leq C \prod_{k=1}^d (t_{k, j_k} - t_{k, j_k - q_k})^{-1/2} \|f|_{I_{j_1, \dots, j_d}}\|_2.$$

PROOF. By equation (12.8) of Theorem 12.2 from [24],

$$\begin{aligned} f(\mathbf{x})|_{I_{j_1, \dots, j_d}} &= \sum_{m_1=1}^{J_1} \cdots \sum_{m_d=1}^{J_d} \theta_{m_1, \dots, m_d} \prod_{k=1}^d B_{m_k, q_k}(x_k)|_{I_{j_1, \dots, j_d}} \\ &= \sum_{l_1=0}^{q_1-1} \cdots \sum_{l_d=0}^{q_d-1} \alpha_{l_1, \dots, l_d} \prod_{k=1}^d x_k^{l_k} \quad \text{for } x_k \in [t_{k, j_k - q_k}, t_{k, j_k}]. \end{aligned}$$

If $\mathbf{x} \in I_{j_1, \dots, j_d}$, then $\mathbf{x} \in \prod_{k=1}^d [t_{k, j_k - h_k - 1}, t_{k, j_k - h_k}]$ for some $h_k = 0, 1, \dots, q_k - 1$. Therefore, this implies that only terms associated with coefficients $\gamma = \{\theta_{m_1, \dots, m_d} : j_k - h_k \leq m_k \leq j_k - h_k + q_k - 1, k = 1, \dots, d\}$ will be nonzero. Furthermore, we define $\alpha = \{\alpha_{l_1, \dots, l_d} : 0 \leq l_k \leq q_k - 1, k = 1, \dots, d\}$. The two equivalent representations of f on I_{j_1, \dots, j_d} above implies a one-to-one mapping between γ and α , i.e., each element of α is a linear combination of elements in γ and vice-versa. Hence, there are matrices \mathbf{T} and \mathbf{V} of dimension $\prod_{k=1}^d q_k \times \prod_{k=1}^d q_k$ respectively, such that $\mathbf{T}\gamma = \alpha$ and $\mathbf{V}\alpha = \gamma$. Since these two linear transformations have entries and dimensions not depending on n , we have $\|\mathbf{T}\|_{(\infty, \infty)} = O(1)$ and $\|\mathbf{V}\|_{(\infty, \infty)} = O(1)$, with constants in $O(1)$ depending only on \mathbf{q} . Let $\mathbf{U}_{k, q_k} = (1, U_k, U_k^2, \dots, U_k^{q_k-1})^T$ where $U_k \sim \text{Uniform}(t_{k, j_k - q_k}, t_{k, j_k})$, $k = 1, \dots, d$. Therefore, $\|f|_{I_{j_1, \dots, j_d}}\|_2^2$ is

$$\begin{aligned} &\int_{I_{j_1, \dots, j_d}} \left(\sum_{l_1=0}^{q_1-1} \cdots \sum_{l_d=0}^{q_d-1} \alpha_{l_1, \dots, l_d} \prod_{k=1}^d x_k^{l_k} \right)^2 d\mathbf{x} \\ &= \sum_{l_1=0}^{q_1-1} \cdots \sum_{l_d=0}^{q_d-1} \sum_{l'_1=0}^{q_1-1} \cdots \sum_{l'_d=0}^{q_d-1} \alpha_{l_1, \dots, l_d} \alpha_{l'_1, \dots, l'_d} \prod_{k=1}^d \int_{[t_{k, j_k - q_k}, t_{k, j_k}]} x_k^{l_k + l'_k} dx_k \\ &\geq \prod_{k=1}^d (t_{k, j_k} - t_{k, j_k - q_k}) \lambda_{\min}\{\mathbf{E}(\mathbf{U}_{k, q_k} \mathbf{U}_{k, q_k}^T)\} \|\alpha\|^2. \end{aligned}$$

Since $\mathbf{E}(\mathbf{U}_{k, q_k} \mathbf{U}_{k, q_k}^T)$ is nonsingular, its minimum eigenvalue is bounded below by a positive constant. Hence, $\lambda_{\min}\{\mathbf{E}(\mathbf{U}_{k, q_k} \mathbf{U}_{k, q_k}^T)\} \|\alpha\|^2 \gtrsim \|\alpha\|_\infty^2 \geq \|\mathbf{V}\|_{(\infty, \infty)}^{-2} \|\gamma\|_\infty^2$. The lower bound is obtained by noting that $\|f|_{I_{j_1, \dots, j_d}}\|_\infty^2 \leq \|\gamma\|_\infty^2 \|\sum_{m_1=1}^{J_1} \cdots \sum_{m_d=1}^{J_d} \prod_{k=1}^d B_{m_k, q_k}(\cdot)|_{I_{j_1, \dots, j_d}}\|_\infty^2 \leq \|\gamma\|_\infty^2$. \square

LEMMA 8.8. For $f(\mathbf{x}) = \mathbf{b}_{\mathbf{J}, \mathbf{q}}(\mathbf{x})^T \boldsymbol{\theta}$, we have

$$\|f\|_2^2 \asymp \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1, \dots, j_d}^2 \prod_{k=1}^d (t_{k, j_k} - t_{k, j_k - q_k}).$$

PROOF. Since $\mathbf{b}_{J,q}(\mathbf{x})$ is a probability vector at any \mathbf{x} , we use Jensen's inequality to write

$$\begin{aligned} \int_{[0,1]^d} f(\mathbf{x})^2 d\mathbf{x} &\leq \int_{[0,1]^d} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1, \dots, j_d}^2 \prod_{k=1}^d B_{j_k, q_k}(x_k) d\mathbf{x} \\ &\leq \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1, \dots, j_d}^2 \int_{[0,1]^d} \prod_{k=1}^d \mathbb{1}_{(t_{k,j_k-q_k}, t_{k,j_k})}(x_k) d\mathbf{x} \\ &= \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1, \dots, j_d}^2 \prod_{k=1}^d (t_{k,j_k} - t_{k,j_k-q_k}). \end{aligned}$$

Using Lemma 8.7 and equation (5) of Chapter XI from [8], $\|f\|_2^2$ is

$$\begin{aligned} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \|f|_{I_{j_1, \dots, j_d}}\|_2^2 &\gtrsim \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \prod_{k=1}^d (t_{k,j_k} - t_{k,j_k-q_k}) \|f|_{I_{j_1, \dots, j_d}}\|_\infty^2 \\ &\geq c \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1, \dots, j_d}^2 \prod_{k=1}^d (t_{k,j_k} - t_{k,j_k-q_k}), \end{aligned}$$

where $c > 0$ is a constant depending only on $\mathbf{q} = (q_1, \dots, q_d)^T$. \square

The following is a multivariate generalization of Lemma 6.1 in [36].

LEMMA 8.9. *For quasi-uniform knots, $\boldsymbol{\theta}^T \mathbf{B}^T \mathbf{B} \boldsymbol{\theta} \asymp n \left(\prod_{k=1}^d J_k^{-1} \right) \|\boldsymbol{\theta}\|^2$ for any $\boldsymbol{\theta} \in \mathbb{R}^J$ if (2.2) holds.*

PROOF. Let $f(\mathbf{x}) = \mathbf{b}_{J,q}(\mathbf{x})^T \boldsymbol{\theta}$ and $\|f\|_{2,\nu}^2 = \int_{[0,1]^d} f(\mathbf{x})^2 d\nu$ for any sigma-finite measure ν . Observe that $\|f\|_{2,G_n}^2 = \boldsymbol{\theta}^T \mathbf{B}^T \mathbf{B} \boldsymbol{\theta} / n$. If the density of G lies between K_{\min} and K_{\max} , then by the quasi-uniformity of the knots and Lemma 8.8, the upper bound for $\|f\|_{2,G}^2$ is

$$(8.7) \quad \|f\|_{2,G}^2 \leq K_{\max} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1, \dots, j_d}^2 \prod_{k=1}^d (t_{k,j_k} - t_{k,j_k-q_k}) \lesssim \|\boldsymbol{\theta}\|^2 \prod_{k=1}^d \Delta_k,$$

and for a constant $c > 0$, the lower bound for $\|f\|_{2,G}^2$ is

$$(8.8) \quad c^2 K_{\min} \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1, \dots, j_d}^2 \prod_{k=1}^d (t_{k,j_k} - t_{k,j_k-q_k}) \gtrsim \|\boldsymbol{\theta}\|^2 \prod_{k=1}^d \Delta_k.$$

Noting that $(G_n - G)(\mathbf{1}_d) = (G_n - G)(\mathbf{0}) = 0$, we use multivariate integration by parts and (2.2) to bound $|\int_{[0,1]^d} f(\mathbf{x})^2 d(G_n - G)(\mathbf{x})|$ by

$$(8.9) \quad \begin{aligned} & 2 \sup_{\mathbf{x} \in [0,1]^d} |G_n(\mathbf{x}) - G(\mathbf{x})| \int_{[0,1]^d} \left| f(\mathbf{x}) \frac{\partial^d f(\mathbf{x})}{\partial x_1 \cdots \partial x_d} \right| d\mathbf{x} \\ & = o \left(\prod_{k=1}^d N_k^{-1} \right) \|f\|_2 \left\| \frac{\partial^d f}{\partial x_1 \cdots \partial x_d} \right\|_2, \end{aligned}$$

in view of the Cauchy-Schwarz inequality in the last line. From (3.1), we have that $\mathfrak{D}_{j_k}^1 \theta_{j_1, \dots, j_k} = (q_k - 1) \mathfrak{d}_{j_k} \theta_{j_1, \dots, j_d} (t_{k, j_k} - t_{k, j_k - q_k + 1})^{-1}$, where $\mathfrak{d}_{j_k} \theta_{j_1, \dots, j_d} = \theta_{j_1, \dots, j_{k-1}, j_{k+1}, j_{k+1}, \dots, j_d} - \theta_{j_1, \dots, j_{k-1}, j_k, j_{k+1}, \dots, j_d}$. Let $\mathfrak{d} \theta_{j_1, \dots, j_d} = \mathfrak{d}_{j_1} \cdots \mathfrak{d}_{j_d} \theta_{j_1, \dots, j_d}$. By setting $\mathbf{r} = \mathbf{1}_d$ in (3.2),

$$\frac{\partial^d f(\mathbf{x})}{\partial x_1 \cdots \partial x_d} = \sum_{j_1=1}^{J_1-1} \cdots \sum_{j_d=1}^{J_d-1} \mathfrak{d} \theta_{j_1, \dots, j_d} \prod_{k=1}^d \frac{q_k - 1}{t_{k, j_k} - t_{k, j_k - q_k + 1}} B_{j_k, q_k - 1}(x_k).$$

Applying Lemma 8.8 to f and its derivatives,

$$\|f\|_2^2 \leq \sum_{j_1=1}^{J_1} \cdots \sum_{j_d=1}^{J_d} \theta_{j_1, \dots, j_d}^2 \prod_{k=1}^d (t_{k, j_k} - t_{k, j_k - q_k}) \lesssim \|\boldsymbol{\theta}\|^2 \prod_{k=1}^d \Delta_k,$$

and $\|\partial^d f / \partial x_1 \cdots \partial x_d\|_2^2$ is bounded by

$$\sum_{j_1=1}^{J_1-1} \cdots \sum_{j_d=1}^{J_d-1} (\mathfrak{d} \theta_{j_1, \dots, j_d})^2 \prod_{k=1}^d \frac{(q_k - 1)^2}{t_{k, j_k} - t_{k, j_k - q_k + 1}} \lesssim \|\boldsymbol{\theta}\|^2 \prod_{k=1}^d \frac{1}{\min_{1 \leq l \leq N_k} \delta_{k, l}},$$

where the last inequality follows from $\sum_{j_1=1}^{J_1-1} \cdots \sum_{j_d=1}^{J_d-1} (\mathfrak{d} \theta_{j_1, \dots, j_d})^2 \leq 2^{2d} \|\boldsymbol{\theta}\|^2$. By the quasi-uniformity of the knots, it follows that the right side of (8.9) is $o(\prod_{k=1}^d N_k^{-1}) \|\boldsymbol{\theta}\|^2$. Combining this result with (8.7) and using Lemma 8.1,

$$\|f\|_{2, G_n}^2 \lesssim \|\boldsymbol{\theta}\|^2 \prod_{k=1}^d \Delta_k + o \left(\prod_{k=1}^d N_k^{-1} \right) \|\boldsymbol{\theta}\|^2 \lesssim \left(\prod_{k=1}^d J_k^{-1} \right) \|\boldsymbol{\theta}\|^2;$$

while combining the same result with (8.8) and in view of Lemma 8.1,

$$\|f\|_{2, G_n}^2 \gtrsim \|\boldsymbol{\theta}\|^2 \prod_{k=1}^d \Delta_k - o \left(\prod_{k=1}^d N_k^{-1} \right) \|\boldsymbol{\theta}\|^2 \gtrsim \left(\prod_{k=1}^d J_k^{-1} \right) \|\boldsymbol{\theta}\|^2. \quad \square$$

LEMMA 8.10. *Let \mathbf{A} be an $n \times n$ symmetric positive definite matrix. Assume that $\|\mathbf{A}\|_{(2,2)} \leq C$ for constant $C > 0$. Let $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$ such that ε_i are i.i.d. mean 0, variance σ_0^2 with finite fourth moment for $i = 1, \dots, n$. Then $\text{Var}(\boldsymbol{\varepsilon}^T \mathbf{A} \boldsymbol{\varepsilon}) = O(n)$.*

PROOF. By eigendecomposition, $\mathbf{A} = \mathbf{P}^T \boldsymbol{\Lambda} \mathbf{P}$ where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \dots, \lambda_n)$ and $\mathbf{P} = ((p_{ij}))$ is an orthogonal matrix. Let $\mathbf{Z} = (Z_1, \dots, Z_n)^T = \mathbf{P} \boldsymbol{\varepsilon}$. Then $\text{Var}(\boldsymbol{\varepsilon}^T \mathbf{A} \boldsymbol{\varepsilon}) = \sum_{i=1}^n \lambda_i^2 \text{Var}(Z_i^2) + \sum_{r \neq s} \lambda_r \lambda_s \text{Cov}(Z_r^2, Z_s^2)$, and

$$\mathbb{E}(Z_r^2 Z_s^2) = \mathbb{E}(\varepsilon_1^4) \sum_{j=1}^n p_{rj}^2 p_{sj}^2 + \sigma_0^4 \sum_{j_1 \neq j_2}^n p_{rj_1}^2 p_{sj_2}^2 + 2\sigma_0^4 \sum_{j_1 \neq j_2}^n p_{rj_1} p_{sj_1} p_{rj_2} p_{sj_2}.$$

Therefore $\sum_{i=1}^n \lambda_i^2 \text{Var}(Z_i^2) \lesssim \sum_{i=1}^n \mathbb{E}(Z_i^4) \lesssim \sum_{i=1}^n (\sum_{j=1}^n p_{ij}^2)^2 \lesssim n$ since $\lambda_i \leq C, i = 1, \dots, n$, and each row of \mathbf{P} has unit norm. By the orthonormality of \mathbf{P} , $\text{Var}(\mathbf{Z}) = \sigma_0^2 \mathbf{I}_n$ and $\mathbb{E}(Z_i^2) = \sigma_0^2$. Observing that for $r \neq s$, $\sum_{j=1}^n p_{rj} p_{sj} = 0$, $\sum_{j_1 \neq j_2}^n p_{rj_1} p_{sj_1} p_{rj_2} p_{sj_2} = (\sum_{j=1}^n p_{rj} p_{sj})^2 - \sum_{j=1}^n p_{rj}^2 p_{sj}^2 \leq 0$ and $\sum_{j_1, j_2=1}^n p_{rj_1}^2 p_{rj_2}^2 = \sum_{j=1}^n p_{rj}^4 + \sum_{j_1 \neq j_2}^n p_{rj_1}^2 p_{rj_2}^2 = 1$. Hence using the last display $\text{Cov}(Z_r^2, Z_s^2)$ is bounded by

$$\mathbb{E}(\varepsilon_1^4) \sum_{j=1}^n p_{sj}^2 p_{rj}^2 + \sigma_0^4 \left(1 - \sum_{j=1}^n p_{rj}^4 \right) - \mathbb{E}(Z_r^2) \mathbb{E}(Z_s^2) \leq \mathbb{E}(\varepsilon_1^4) \sum_{j=1}^n p_{sj}^2 p_{rj}^2.$$

Therefore, $\sum_{r \neq s} \lambda_r \lambda_s \text{Cov}(Z_r^2, Z_s^2) \lesssim \mathbb{E}(\varepsilon_1^4) \sum_{j=1}^n \sum_{r=1}^n \sum_{s=1}^n p_{sj}^2 p_{rj}^2 \lesssim n$. \square

LEMMA 8.11. *Let $\{X(\mathbf{t}) : \mathbf{t} \in [0, 1]^d\}$ be a sub-Gaussian process with respect to the semi-metric $d(\mathbf{t}, \mathbf{s}) = \sqrt{\text{Var}[X(\mathbf{t}) - X(\mathbf{s})]}$ such that $d^2(\mathbf{t}, \mathbf{s}) \lesssim C(n) \|\mathbf{t} - \mathbf{s}\|^2$ for any $\mathbf{t}, \mathbf{s} \in [0, 1]^d$, where $C(n)$ is a polynomial in n . Choose points $\mathbf{u} = \{\mathbf{u}_1, \dots, \mathbf{u}_{T_n}\}$ in $[0, 1]^d$ such that $\bigcup_{i=1}^{T_n} \{\mathbf{z} : \|\mathbf{z} - \mathbf{u}_i\| \leq \delta_n\} \supseteq [0, 1]^d$, for some sequence $\delta_n \rightarrow 0$ as $n \rightarrow \infty$ with $\delta_n < 1$, and $T_n \leq (2/\delta_n)^d$. Then for $1 \leq p < \infty$,*

$$\begin{aligned} \mathbb{E} \|X\|_\infty^p &\lesssim \{\log(1/\delta_n)\}^{p/2} \left\{ \left(\delta_n \sqrt{C(n)} \right)^p \right. \\ &\quad \left. + \max_{1 \leq i \leq T_n} |\mathbb{E}[X(\mathbf{u}_i)]|^p + \max_{1 \leq i \leq T_n} \{\text{Var}[X(\mathbf{u}_i)]\}^{p/2} \right\}. \end{aligned}$$

PROOF. It suffices to bound the L_p -norms of the expected process increment and the maximum of the process at \mathbf{u} . Since $X(\mathbf{t})$ is sub-Gaussian and $\text{Var}[X(\mathbf{t}) - X(\mathbf{s})] \lesssim C(n) \|\mathbf{t} - \mathbf{s}\|^2$ by assumption, we can relate the ψ_2 -Orlicz norm of the process increment with $\|\mathbf{t} - \mathbf{s}\|$ by Section 2.2.1 of [35]. We then bound the expected process increment by Corollary 2.2.8 of

[35] with $d(\mathbf{t}, \mathbf{s}) = \|\mathbf{t} - \mathbf{s}\|$. The expected maximum of the process at \mathbf{u} is then bounded using Lemma 2.2.2 of [35]. \square

Acknowledgment. The authors like to thank the referees and the associate editors for suggestions which led to significant improvements of the paper and for directing to several relevant references.

References.

- [1] BICKEL, P. and ROSENBLATT, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **1** 1071–1095. Correction (1975) **3** 1370.
- [2] CASTILLO, I. (2014). On Bayesian supremum norm contraction rates. *Ann. Statist.* **42** 2058–2091.
- [3] CASTILLO, I. and NICKL, R. (2013). Nonparametric Bernstein-von Mises theorems in Gaussian white noise. *Ann. Statist.* **41** 1999–2028.
- [4] CASTILLO, I. and NICKL, R. (2014). On the Bernstein-von Mises phenomenon for nonparametric Bayes procedures. *Ann. Statist.* **42** 1941–1969.
- [5] CHERNOZHUKOV, V., CHETVERIKOV, D. and KATO, K. (2014). Anti-concentration and honest, adaptive confidence bands. *Ann. Statist.* **42** 1787–1818.
- [6] CLAESKENS, G. and VAN KEILEGOM, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *Ann. Statist.* **31** 1852–1884.
- [7] COX, D. D. (1993). An analysis of Bayesian inference for nonparametric regression. *Ann. Statist.* **21** 903–923.
- [8] DE BOOR, C. (2001). *A Practical Guide to Splines*, Revised ed. Springer-Verlag New York, Inc.
- [9] DE JONGE, R. and VAN ZANTEN, J. H. (2012). Adaptive estimation of multivariate functions using conditionally Gaussian tensor-product spline priors. *Electron. J. Stat.* **6** 1984–2001.
- [10] DE JONGE, R. and VAN ZANTEN, J. H. (2013). Semiparametric Bernstein-von Mises for the error standard deviation. *Electron. J. Stat.* **7** 217–243.
- [11] DEMKO, S., MOSS, W. F. and SMITH, P. W. (1984). Decay rates for inverses of band matrices. *Math. Comp.* **43** 491–499.
- [12] FREEDMAN, D. (1999). On the Bernstein-von Mises theorem with infinite dimensional parameters. *Ann. Statist.* **27** 1119–1140.
- [13] GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170.
- [14] GINÉ, E. and NICKL, R. (2011). Rates of contraction for posterior distributions in L^r -metrics, $1 \leq r \leq \infty$. *Ann. Statist.* **39** 2883–2911.
- [15] HARVILLE, D. A. (1997). *Matrix Algebra from a Statistician's Perspective*. Springer-Verlag New York, Inc.
- [16] HOFFMANN, M. and LEPSKI, O. (2002). Random rates in anisotropic regression. *Ann. Statist.* **30** 325–396.
- [17] HOFFMANN, M., ROUSSEAU, J. and SCHMIDT-HIEBER, J. (2015). On adaptive posterior concentration rates. *Ann. Statist.* To appear.
- [18] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2011). Bayesian inverse problems with Gaussian priors. *Ann. Statist.* **39** 2626–2657.
- [19] KNAPIK, B. T., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2013). Bayesian recovery of the initial condition for the heat equation. *Comm. Statist. Theory Methods* **42** 1294–1313.

- [20] LEAHU, H. (2011). On the Bernstein-von Mises phenomenon in the Gaussian white noise model. *Electron. J. Stat.* **5** 373–404.
- [21] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*. Springer-Verlag Berlin Heidelberg.
- [22] NEUMANN, M. H. and VON SACHS, R. (1997). Wavelet thresholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.* **25** 38–76.
- [23] RAY, K. (2015). Adaptive Bernstein-von Mises theorems in Gaussian white noise. Preprint arXiv:1407.3397v2 [math.ST].
- [24] SCHUMAKER, L. (2007). *Spline Functions: Basic Theory*, Third ed. Cambridge University Press, New York.
- [25] SCRICCILO, C. (2014). Adaptive Bayesian density estimation in L^p -metrics with Pitman-Yor or normalized inverse-Gaussian process kernel mixtures. *Bayesian Analysis* **9** 475–520.
- [26] SEARLE, S. R. (1982). *Matrix Algebra Useful for Statistics*. John Wiley and Sons, Inc.
- [27] SERRA, P. and KRIVOBOKOVA, T. (2014). Adaptive empirical Bayesian smoothing splines. Preprint arXiv:1411.6860 [math.ST].
- [28] SHEN, W. and GHOSAL, S. (2014). Adaptive Bayesian procedures using random series priors. *Scand. J. Statist.* To appear.
- [29] SHEN, W. and GHOSAL, S. (2014). Adaptive Bayesian density regression for high-dimensional data. *Bernoulli*. To appear.
- [30] SMIRNOV, N. V. (1950). On the construction of confidence regions for the density of distribution of random variables. *Doklady Akad. Nauk SSSR* **74** 189–191.
- [31] SNEKERS, S. and VAN DER VAART, A. W. (2015). Credible sets in the fixed design model with Brownian motion prior. *J. Statist. Plann. Inference* **166** 78–86.
- [32] STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.
- [33] STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.
- [34] SZABO, B., VAN DER VAART, A. W. and VAN ZANTEN, J. H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428.
- [35] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Process With Applications to Statistics*. Springer-Verlag New York, Inc.
- [36] ZHOU, S., SHEN, X. and WOLFE, D. A. (1998). Local Asymptotics for Regression Splines and Confidence Regions. *Ann. Statist.* **26** 1760–1782.

W. W. Yoo
 UNIVERSITÉ PARIS DAUPHINE
 CEREMADE
 PLACE DU MARÉCHAL DE LATTRE DE TASSIGNY,
 75016 PARIS, FRANCE
 E-MAIL: yoowein0203@gmail.com

S. GHOSAL
 DEPARTMENT OF STATISTICS
 NORTH CAROLINA STATE UNIVERSITY
 4276 SAS HALL, 2311 STINSON DRIVE
 RALEIGH, NORTH CAROLINA 27695-8203
 USA
 E-MAIL: sghosal@ncsu.edu